KLASIFIKASI FILM BERDASARKAN SINOPSIS DENGAN MENGGUNAKAN IMPROVED K-NEAREST NEIGHBOR (K-NN)

SKRIPSI

Untuk memenuhi sebagian persyaratan
memperoleh gelar Sarjana Komputer

Disusun oleh:
Nurul Muslimah
NIM: 145150201111139

PROGRAM STUDI TEKNIK INFORMATIKA
JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS BRAWIJAYA
MALANG
2018

PENGESAHAN

KLASIFIKASI FILM BERDASARKAN SINOPSIS DENGAN MENGGUNAKAN IMPROVED K-NEAREST NEIGHBOR (K-NN)

SKRIPSI

Diajukan untuk memenuhi sebagian persyaratan
memperoleh gelar Sarjana Komputer

Disusun Oleh :
Nurul Muslimah
NIM: 145150201111139

Skripsi ini telah diuji dan dinyatakan lulus pada
2 Agustus 2018
Telah diperiksa dan disetujui oleh:

Dosen Pembimbing I	Dosen Pembimbing II
Indriati, S.T, M.Kom	Randy Cahya Wihandika, S.ST ., M.Kom
NIP: 19831013 201504 2 002	NIK: 201405 880206 1 001

Mengetahui
Ketua Jurusan Teknik Informatika

Tri Astoto Kurniawan, S.T, M.T, Ph.D
NIP: 19710518 200312 1 001

PERNYATAAN ORISINALITAS

Saya menyatakan dengan sebenar-benarnya bahwa sepanjang pengetahuan saya, di dalam naskah skripsi ini tidak terdapat karya ilmiah yang pernah diajukan oleh orang lain untuk memperoleh gelar akademik di suatu perguruan tinggi, dan tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis disitasi dalam naskah ini dan disebutkan dalam daftar pustaka.

Apabila ternyata didalam naskah skripsi ini dapat dibuktikan terdapat unsur-unsur plagiasi, saya bersedia skripsi ini digugurkan dan gelar akademik yang telah saya peroleh (sarjana) dibatalkan, serta diproses sesuai dengan peraturan perundang-undangan yang berlaku (UU No. 20 Tahun 2003, Pasal 25 ayat 2 dan Pasal 70).

Malang, 19 Juli 2018

Nurul Muslimah

NIM: 145150201111139

KATA PENGANTAR

Puji syukur atas kehadirat Allah SWT atas segala karunianya yang telah melimpahkan rahmat, taufik, dan hidayah-Nya sehingga laporan penelitian skripsi ini yang berjudul "Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-Nearest Neighbor (K-NN)" dapat terselesaikan dengan baik.

Melalui kesempatan ini, penulis menyadari penulisan skripsi ini tidak akan dapat terselesaikan jika tanpa bantuan dari berbagai pihak. Oleh sebab itu, penulis ingin menyampaikan rasa terimakasih dan hormat yang sebesar-besarnya kepada segala pihak yang telah mendukung, memberikan bantuan, serta doa selama proses penulisan skripsi, diantaranya:

1. Ibu Indriati, S.T, M.Kom dan Bapak Randy Cahya Wihandika, S.ST., M.Kom selaku dosen pembimbing skripsi yang telah membimbing dan mengarahkan penulis dengan sabar sehingga penelitian skripsi ini dapat terselesaikan
2. Bapak Wayan Firdaus Mahmudy, S.Si, M.T, Ph.D., Bapak Ir. Heru Nurwarsito, M.Kom, Bapak Drs. Marji, M.T, dan Bapak Edy Santoso, S.Si, M.Kom selaku Dekan, Wakil Dekan I, Wakil Dekan II dan Wakil Dekan III Fakultas Ilmu Komputer Universitas Brawijaya.
3. Bapak Tri Astoto Kurniawan, S.T, M.T, Ph.D, Bapak Agus Wahyu Widodo, S.T, M.Cs dan Bapak Muhammad Tanzil Furqon, S.Kom, M.CompSc selaku Ketua Jurusan, Ketua Program Studi dan Sekretaris Program Studi Teknik Informatika.
4. Ayahanda Abdul Karim M. Ali dan Ibunda Mariati S.Pd yang telah memberikan motivasi, dukungan, kasih sayang, perhatian, serta senantiasa tiada hentinya memberikan doa demi kelancaran dan terselesaikannya skripsi ini.
5. Saudara-saudara saya, Rakhmat Hidayat, Sri Rahayu, Ratih Kurniati, dan Nur Amalia, serta seluruh keluarga besar tercinta yang selalu mendukung dan memberikan doa demi kelancaran skripsi ini.
6. Seluruh civitas akademika Teknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya yang telah memberikan bantuan dan dukungan selama penulis menempuh studi dan selama penyelesaian skripsi di Teknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya.
7. Teman-teman terdekat saya, Putu Amelia Vennanda W., Chandra Ayu A. P., Riska Dewi Nurfarida, dan Nana Nofiana yang telah membantu dan memberikan dukungan selama proses penyelesaian penelitian skripsi ini.

8. Teman-teman Teknik Informatika angkatan 2014 dan Komputasi Cerdas serta seluruh pihak yang telah membantu kelancaran penulisan skripsi yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa penyusunan skripsi ini masih memiliki banyak kekurangan, sehingga penulis membutuhkan adanya kritik maupun saran yang bersifat membangun. Akhir kata dari penulis, saya harap skripsi ini dapat memberikan manfaat bagi semua pihak yang menggunakannya.

Malang, 19 Juli 2018

Penulis

nurulmuslimah25@gmail.com

ABSTRAK

Film merupakan media komunikasi bersifat audio visual, dimana tersirat pesan yang ingin disampaikan pencipta film. Film memiliki beberapa genre yakni romantis, horor, *thriller*, komedi, fantasi dan lain sebagainya. Tidak sedikit penikmat film yang masih bingung akan perbedaan dari genre-genre tersebut. Hal tersebut mengakibatkan banyaknya penikmat film yang susah untuk membedakan genre film sehingga pesan pada film tak sepenuhnya dapat tersampaikan kepada penikmat film. Oleh sebab itu melakukan klasifikasi pada film berdasarkan sinopsis film dirasa dapat menjadi salah satu solusi untuk masalah tersebut. Pengklasifikasian pada sinopsis film akan membantu dalam mengelompokan film dengan genre yang sesuai. Proses klasifikasi genre film berdasarkan sinopsis dimulai dengan melakukan *preprocessing*, kemudian pembobotan *term* hingga klasifikasi dengan metode Improved K-NN. Berdasarkan implementasi serta pengujian yang dilakukan pada penelitian Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN yang mana menggunakan 250 dokumen sebagai data latih dan 50 dokumen sebagai data uji didapatkan hasil terbaik yakni *precision* sebesar 1, *recall* sebesar 0,88, *f-measure* sebesar 0,936170213, dan tingkat akurasi sebesar 88%. Selain itu dilakukan perbandingan dengan metode K-NN dan terbukti bahwa pengklasifikasian dengan menggunakan metode Improved K-NN lebih baik dibandingkan dengan metode K-NN.

Kata kunci : *Text Mining*, Klasifikasi, Film, Sinopsis, Improved K-Nearest Neighbor.

ABSTRACT

Movies are audio visual communication media, which imply the message that the movie creator wants to convey. Movie has several genres namely romantic, horror, thriller, comedy, fantasy and so on. Not a few movie connoisseurs are still confused about the differences in these genres. This resulted in many movie lovers who were difficult to distinguish the genre of movie so that the message in the movie could not be fully conveyed to the audience of the movie. Therefore, the classification of movies based on the synopsis of the movie can be one of the solutions to the problem. Classification in the movie synopsis will help in grouping movies with the appropriate genre. The genre genre classification process based on the synopsis begins with preprocessing, then weighting the term to classification with the Improved K-NN method. Based on the implementation and testing conducted on the movie classification research based on the synopsis by using Improved K-NN which uses 250 documents as training data and 50 documents as the test data obtained the best results namely precision by 1, recall by 0.88, f-measure by 0.936170213, and an accuracy rate of 88%. In addition, the K-NN method was compared and it was proved that classification using the Improved K-NN method was better than the K-NN method.

Keywords: Text Mining, Classification, Movie, Synopsis, Improved K-Nearest Neighbor

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Film merupakan salah satu media untuk berkomunikasi yang memiliki sifat audio visual dimana tersirat pesan yang ingin disampaikan oleh pencipta film. Pesan pada film dapat berbeda tergantung dari genre film, umumnya pesan yang disampaikan oleh film berupa pesan tentang pendidikan, informasi, serta untuk hiburan. Saat ini film memiliki banyak ragam yang dapat menjadi pilihan bagi penikmatnya, selain bertujuan untuk menghibur atau memberikan informasi, film juga dapat diciptakan untuk memberikan pelayanan keperluan pada publik.

Secara umum film dapat digolongkan menjadi dua golongan yaitu film fiksi dan non fiksi. Film fiksi diciptakan berdasarkan kisah yang tidak nyata atau dikarang, sedangkan film non fiksi diciptakan berdasarkan kisah nyata. Dari dua golongan film tersebut dilahirkanlah beberapa genre film yang beragam yakni romantis, horor, *thriller*, komedi, fantasi dan lain sebagainya. Dari banyaknya genre yang disediakan oleh film, tak sedikit penikmat film yang masih bingung akan perbedaan dari genre-genre tersebut. Hal tersebut mengakibatkan banyaknya penikmat film yang susah untuk membedakan film sehingga pesan pada film yang ingin disampaikan tak sepenuhnya dapat diterima oleh penikmat. Oleh karena itu melakukan Klasifikasi pada film dirasa penting untuk mempermudah penikmat atau penonton dalam memilih genre yang tepat dan sesuai dengan yang diinginkan.

Klasifikasi ialah pengelompokan sesuatu berdasarkan karateristiknya ke dalam kelas-kelas yang berbeda. Klasifikasi teks adalah mengelompokan dokumen atau teks ke dalam kelas-kelas yang memiliki kemiripan karateristik yang sama atau mirip. Klasifikasi memungkinkan untuk mengelompokan film ke dalam kelas yang sesuai berdasarkan kategorinya. Tidak sedikit penikmat film yang masih bingung membedakan atau menentukan genre film yang sesuai dengan yang diinginkan, serta agar pesan pada film dapat ditujukan dan disampaikan dengan tepat maka melakukan pengelompokan film atau klasifikasi pada sinopsis film dirasa menjadi solusi yang tepat untuk masalah tersebut. Melakukan klasifikasi pada sinopsis film akan membantu dalam mengelompokan film dengan genre yang sesuai. Metode untuk melakukan klasifikasi dokumen tidaklah sedikit yakni KNN, K-Means, Naïve Bayes, SVM dan lain sebagianya. Penelitian ini menggunakan metode Improved K-NN (K-Nearest Neighbor).

Metode K-NN ialah metode untuk melakukan pengelompokan objek sesuai dengan jarak yang paling dekat dari objek dengan masing-masing kategori (Sreemathy dan Balamurugan, 2012). Namun dalam penerapannya metode K-NN memiliki kekurangan yaitu ketika melakukan penentuan kelas dari data kandidat hasil yang didapat masih kurang tepat, maka dengan menggunakan metode Improved K-NN dapat menjadi solusi yang tepat untuk mengatasi masalah tersebut (Megantara et all, 2010). Perbedaan antara metode K-NN dan Improved K-NN terdapat pada penentuan nilai *k*, pada K-NN nilai *k* yang ditentukan pada

tiap kategori ialah memiliki nilai yang sama, sedangkan pada Improved K-NN digunakan nilai *k* yang berbeda pada tiap kategori yang sesuai dengan banyaknya data latih (Puspitasari et all, 2017). Sehingga nilai akurasi yang didapatkan akan lebih tinggi dan maksimal. Metode ini dirasa tepat untuk melakukan klasifikasi sehingga dapat menghasilkan kelas-kelas yang sesuai.

Puspitasari dkk. (2017) pada penelitiannya melakukan pembahasan mengenai pengklasifikasian yang dilakukan pada dokumen tumbuhan obat dengan menggunakan metode Improved K-NN dimana didapatkan perolehan *F1-measure* sebanyak 70,99%, dan dari pengujian data latih ditemukan bahwa semakin besar jumlah data latih maka nilai akurasi akan semakin tinggi, serta diperoleh *F1-measure* untuk data latih seimbang sebesar 1,9% lebih baik dari data latih tidak seimbang. Selain itu penelitian lainnya ialah penelitian yang dilakukan oleh Megantara dkk. (2010) membahas tentang klasifikasi teks dengan menggunakan Improved K-NN, dari hasil penelitian terbukti bahwa metode Improved K-NN mencapai hasil yang lebih baik dibandikan metode K-NN dalam berbagai kondisi, dan dilihat dari standar deviasi yang didapatkan metode Improved K-NN memiliki kestabilan yang lebih baik dari K-NN. Pada penelitian yang dilakukan oleh Nathania dkk. (2017) yang membahas tentang melakukan klasifikasi spam pada twitter menggunakan metode Improved K-NN, didapatkan hasil rata-rata nilai *Precision* sebesar 0,8946 dan *Recall* sebesar 0,9405 serta *F-Measure* sebesar 0,9155. Hasil akurasi diperoleh sebesar 89,57%.

Untuk mendapatkan Kelas-kelas yang sesuai dan memiliki tingkat akurasi yang tepat diperlukan dilakukannya beberapa proses. Proses-proses tersebut akan dibahas lebih jelas dalam skripsi ini.

1.2 Rumusan Masalah

Berdasar dari uraian di atas, maka dapat dirumuskan permasalahan permasalahan yang ada pada skripsi berikut ialah :

1. Bagaimana proses dalam melakukan klasifikasi film berdasarkan sinopsis dengan metode Improved K-NN ?
2. Bagaimana hasil penerapan metode Klasifikasi Improved K-NN pada klasifikasi film berdasarkan sinopsis ?
3. Bagaimana perbandingan hasil menggunakan Improved K-NN dengan KNN ?

1.3 Tujuan

Adapun tujuan dari penelitian Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN ialah sebagai berikut :

1. Mengetahui proses dalam melakukan klasifikasi film berdasarkan sinopsis dengan metode Improved K-NN.
2. Mengetahui hasil penerapan metode Klasifikasi Improved K-NN untuk melakukan klasifikasi film berdasarkan sinopsis.

3. Menguji hasil klasifikasi dan membandingkan hasil klasifikasi yang menggunakan Improved K-NN dengan K-NN.

1.4 Manfaat

Penelitian ini diharapkan memiliki manfaat yang baik serta berguna bagi pembaca dan penulis. Adapun manfaat yang dimiliki adalah sebagai berikut :

Bagi Penulis

1. Sebagai media untuk pengimplementasian ilmu pengetahuan teknologi dalam bidang Information Retrieval terutama Klasifikasi Text.
2. Mendapatkan pengetahuan dan wawasan terkait metode-metode yang digunakan dalam melakukan Klasifikasi.

Bagi pembaca

1. Mendapatkan wawasan akan pengimplementasian dari Improved K-NN dalam Klasifikasi Text.
2. Membantu dalam penentuan Klasifikasi terhadap film.

1.5 Batasan Masalah

Dalam penelitian Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN ini batasan penelitan yang digunakan adalah:

1. Data yang digunakan merupakan sinopsis film berbahasa Indonesia.
2. Data yang digunakan berasal dari sinopsisfilem21.com, posfilm.com, filmbioskop.co.id, pusatsinopsis.com, filmbioskop.net, hype.idntimes.com, filmbor.com, sinopsisfilm.co.id, sinopsisdanreviewfilm.blogspot.com, dan industry.co.id.
3. Metode dalam melakukan pengklasifikasian ialah Improved K-NN.

1.6 Sistematika Pembahasan

Adapun sistematika penulisan dalam skripsi ini ialah sebagai berikut :

BAB I Pendahuluan

Bab ini berisi tentang latar belakang, rumusan masalah, tujuan, batasan masalah, manfaat, dan sistematika penulisan dalam penelitian Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN.

BAB II Tinjauan Pustaka

Tinjauan pustaka menjelaskan tentang kajian pustaka terkait dengan penelitian Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN.

BAB III Metodologi Penelitian

Metodologi Penelitian menjelaskan tentang metode yang digunakan dalam penelitian Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN.

BAB IV Perancangan dan Implementasi

Perancangan menjelaskan tentang analisis kebutuhan serta perancangannya, yaitu aplikasi Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN. Implementasi menjelaskan tentang pengimplementasian dari metode yang digunakan yaitu Improved K-NN pada Klasifikasi Film Berdasarkan Sinopsis.

BAB V Pengujian dan Analisis

Pengujian dan analisis menjelaskan tentang suatu proses dengan hasil pengujian pada Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN serta memberikan analisis pada hasil pengujian yang dilakukan.

BAB VI Penutup

Penutup berisi kesimpulan yang telah diperoleh dari perancangan, implementasi, dan pengujian pembuatan serta saran-saran untuk pengembangan sistem lebih lanjut pada Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN.

BAB 2 LANDASAN KEPUSTAKAAN

Pada bab landasan kepustakaan berikut dibahas mengenai landasan kepustakaan yang digunakan pada penelitian Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN.

2.1 Kajian Pustaka

Puspitasari dkk. (2017) pada penelitiannya melakukan pembahasan mengenai pengklasifikasian yang dilakukan pada dokumen tumbuhan obat dengan menggunakan metode Improved K-NN dimana didapatkan perolehan *F1-measure* sebanyak 70,99%, dan dari pengujian data latih ditemukan bahwa semakin besar jumlah data latih maka nilai akurasi akan semakin tinggi, serta diperoleh *F1-measure* untuk data latih seimbang sebesar 1,9% lebih baik dari data latih tidak seimbang.

Selain itu Megantara dkk. (2010) dalam penelitiannya membahas tentang klasifikasi teks dengan menggunakan Improved K-NN, dari hasil penelitian terbukti bahwa metode Improved K-NN mencapai hasil yang lebih baik dibandingkan metode K-NN dalam berbagai kondisi, dan dilihat dari standar deviasi yang didapatkan metode Improved K-NN memiliki kestabilan yang lebih baik dari K-NN.

Pada penelitian yang dilakukan oleh Nathania dkk. (2017) yang membahas tentang melakukan klasifikasi *spam* pada twitter menggunakan metode Improved K-NN, didapatkan hasil rata-rata nilai *Precision* sebesar 0,8946 dan *Recall* sebesar 0,9405 serta *F-Measure* sebesar 0,9155. Hasil akurasi diperoleh sebesar 89,57%.

Dalam penelitian ini metode yang digunakan ialah metode Improved K-NN, metode tersebut dirasa sesuai untuk melakukan klasifikasi. Diharapkan dengan menggunakan metode tersebut dapat menghasilkan klasifikasi yang sesuai serta memiliki tingkat akurasi yang maksimal.

2.2 Film

Film merupakan salah satu karya seni yang menggabungkan suara dan gambar, dalam film pula terdapat unsur-unsur seni yang menunjang film seperti seni fotografi, tari, puisi,teater,musik dan lain sebagainya. Film juga merupakan salah satu media untuk berkomunikasi yang memiliki sifat audio visual dimana tersirat pesan yang ingin disampaikan ole pencipta film. Pesan pada film dapat berbeda tergantung dari genre film, umumnya pesan yang disampaikan oleh film berupa pesan tentang pendidikan, informasi, serta untuk hiburan. Saat ini film memiliki banyak ragam yang dapat menjadi pilihan bagi penikmatnya, selain bertujuan untuk menghibur atau memberikan informasi, film juga dapat diciptakan untuk memberikan pelayanan keperluan pada publik.

Selain untuk menghibur, film sendiri memiliki banyak fungsi lain yaitu untuk memberikan informasi, pendidikan, serta persuasif. Pada film pesan yang disampaikan disesuaikan dengan fungsi film yang disajikan, sehingga pesan yang disampaikan bisa tersampaikan lebih baik.

Film memiliki beberapa genre yakni horor, romantis, komedi, *thriller*, fantasi, drama dan aksi.

1. Horor
   Genre film ini menghadirkan suasana yang menakutkan, dimana konflik bukan hanya terjadi secara mental melainkan fisik dan emosi pun ikut terlibat, elemen ketakutan pun dibuat lebih menonjol sehingga kesan menakutkan nampak jelas. Cerita yang diangkatpun tak jauh dari hal-hal yang berbau mistis, kematian, supranatural, dan hal-hal yang tak masuk akal.
2. Romantis
   Pada film dengan genre ini akan menceritakan hal yang berkaitan dengan percintaan, dan konflik yang diciptakan pun konflik yang terjadi antar pasangan.
3. Komedi
   Pada film dengan genre ini lebih ditekankan pada unsur komedi, dimana film dengan genre ini menghadirkan suasana yang akan membuat penikmatnya tertawa.
4. *Thriller*
   Pada film dengan genre ini akan menghadirkan suasana yang menegangkan, selain itu tema cerita yang diangkatpun akan membahas tentang misteri, mata-mata, dan konspirasi.
5. Fantasi
   Pada film fantasi tokoh dan cerita yang diangkat merupakan hal-hal yang tidak nyata. Masa lampau atau masa depan merupakan latar waktu yang digunakan untuk genre film ini.
6. Aksi (Action)
   Pada film dengan genre ini menghadirkan cerita dengan lebih menekankan pada aksi dan kekerasan dan memiliki tokoh antagonis yang memiliki peran yang jelas, serta menghadirkan tokoh protagonist sebagai pahlawan utama yang akan menyelesaikan masalah.

2.3 Text Mining

Text mining merupakan proses melakukan penambangan data dari dokumen atau data-data yang tidak terstruktur. *Text mining* berusaha untuk melakukan pengekstrakan informasi yang tersirat secara implisit dari informasi yang telah diekstrak secara otomatis dari sumber data atau dokumen (Feldman., 2007).

Text mining memiliki tujuan untuk mendapakatkan hasil berupa informasi yang berguna dari kumpulan-kumpulan dokumen, adapun sumber data yang digunakan untuk melakukan *text mining* ialah berupa data yang tidak terstruktur

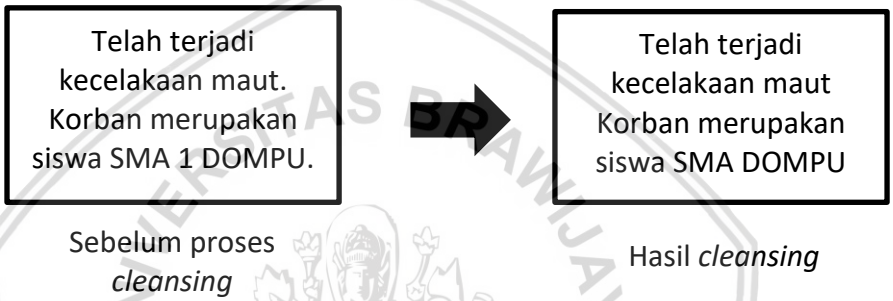
atau semi terstruktur. Beberapa hal yang bisa dilakukan dengan menggunakan *text mining* ialah melakukan Klasifikasi *text*, *Clustering text*, dan lain sebagainya.

2.4 Text Preprocessing

Text Preprocessing merupakan tahap awal dalam tahapan *text mining*, pada tahap ini data-data yang belum terstruktur dibersihkan agar menjadi lebih terstruktur, terdapat beberapa proses untuk melakukan *Preprocessing* yakni *cleansing, case folding*, tokenisasi*, filtering*, dan *stemming*.

1. *Cleansing*

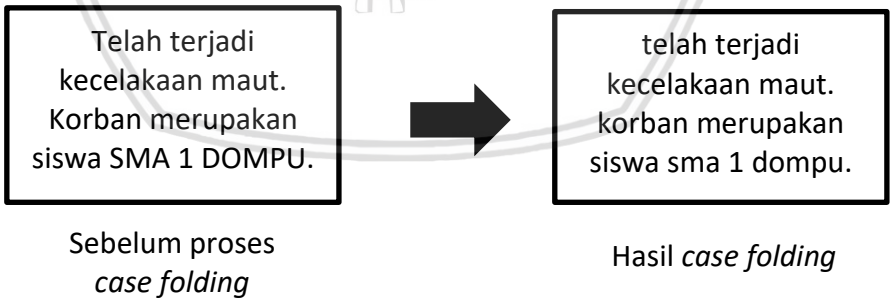
   Cleansing merupakan proses untuk membersihkan *text* dari karakter-karakter yang tidak perlu serta menghapus *link* pada dokumen. Berikut contoh proses *cleansing* dapat dilihat pada Gambar 2.1



Gambar 2.1 Contoh proses *cleansing*

2. *Case Folding*

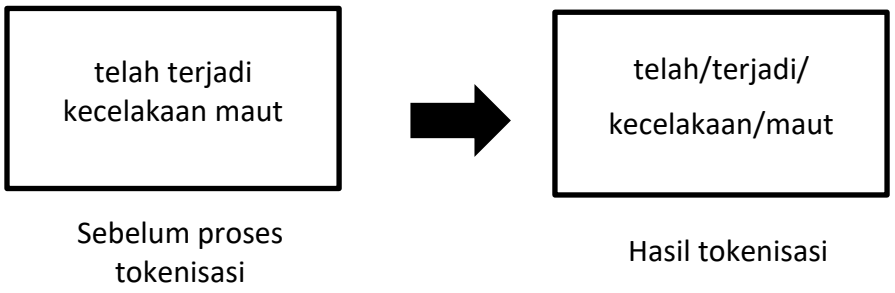
   Case folding merupakan proses yang dilakukan untuk mengubah seluruh huruf yang terdapat pada dokumen atau data menjadi huruf kecil. Berikut contoh proses *case folding* dapat dilihat pada Gambar 2.2



Gambar 2.2 Contoh proses *case folding*

3. Tokenisasi

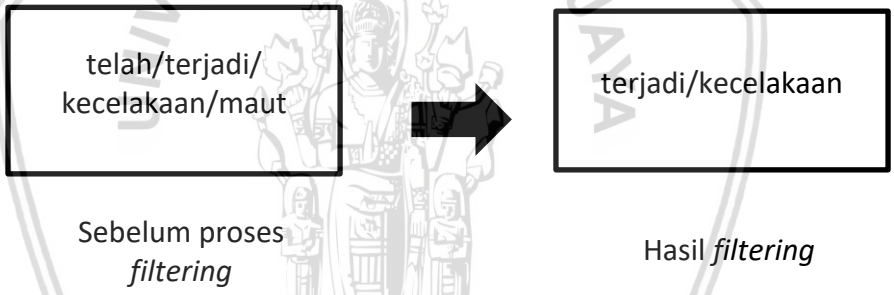
   Tokenisasi ialah proses untuk melakukan pemotongan pada string-string yang terdapat dalam dokumen. Adapun contoh proses tokenisasi dapat dilihat pada Gambar 2.3



Gambar 2.3 Contoh proses tokenisasi

4. *Filtering*

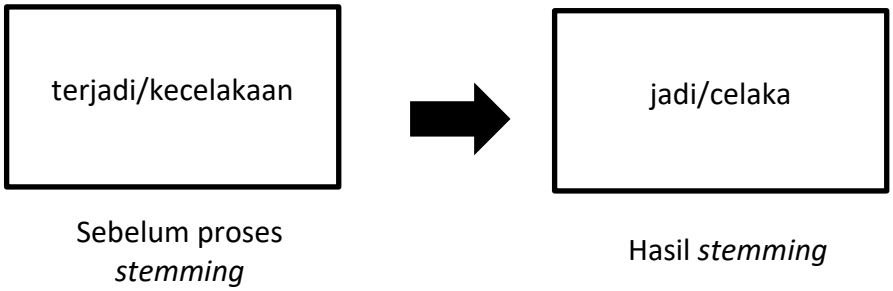
Filtering merupakan proses mengambil kata inti atau kata utama dari dokumen, serta menghilangkan kata yang tidak memiliki makna atau kata yang tidak diperlukan. Proses *filtering* dapat menggunakan *stopword* yaitu kata yang sering muncul dalam dokumen namun tidak memiliki arti ataupun makna sehingga dirasa tidak penting. Serta proses ini dapat menggunakan *wordlist* ialah sekumpulan kata yang memiliki makna dan merupakan kata penting. Contoh proses *filtering* dapat dilihat pada Gambar 2.4



Gambar 2.4 Contoh proses *filtering*

5. *Stemming*

Stemming merupakan proses menghilangkan imbuhan pada suatu kata sehingga hanya tersisa kata akarnya saja. Adapun contoh proses stemming dapat dilihat pada Gambar 2.5



Gambar 2.5 Contoh proses *stemming*

2.4.1 Algoritma *Stemming* Nazief dan Andriani

Algoritma *stemming* Nazief dan Andriani ialah algoritma yang dikembangkan sesuai dengan aturan susunan Bahasa Indonesia yang melakukan pengelompokan imbuhan menjadi prefix (awalan), infix (sisipan), suffix (akhiran), serta gabungan antara prefix dan infix yakni confixes. (Wahyudi dkk, 2013). Dalam melakukan *stemming* ada beberapa langkah-langkah yang harus dilakukan, langkah-langkah tersebut ialah sebagai berikut :

1. Melakukan pencarian kata yang belum melewati proses *stemming* pada kamus, jika kata terdapat pada kamus maka kata tersebut merupakan kata yang tepat maka proses algoritma akan dihentikan.
2. Menghilangkan *inflectional endings* yaitu melakukan penghapusan pada kumpulan huruf yang diletakan diakhir kata yang memiliki fungsi untuk mengubah makna dari kata seperti ‘’-lah”, ”-kan”, “-pun”, atapun “-kah”. Setelah itu melakukan penghapusan pada *inflectional possessive pronoun suffixes* seperti “-ku”, “-mu”, dan “nya”. Kemudian melakukan pengecekan terhadap kata dasar pada kamus, apabila kata ditemukan maka proses algoritma akan dihentikan dan apabila tidak ditemukan kata pada kamus maka akan dilanjutkan kelangkah berikutnya.
3. Melakukan penghapusan *derivational suffix* yakni “-i” ataupun “-an”. Kemudian melakukan pengecekan terhadap kata dasar pada kamus, apabila kata ditemukan maka proses algoritma akan dihentikan dan apabila tidak ditemukan kata pada kamus maka akan dilanjutkan kelangkah 3a.
    a. Apabila akhiran berupa “-an” telah dihilangan serta huruf terakhir pada kata ialah “-k”, maka “-k” akan dihilangkan. Kemudian melakukan pengecekan terhadap kata dasar pada kamus, apabila kata ditemukan maka proses algoritma akan dihentikan dan apabila tidak ditemukan kata pada kamus maka akan dilanjutkan kelangkah 3b.
    b. Akhiran kata yang dihilangkan yakni “-i”, “-an” ataupun “-kan” akan dikembalikan, kemudian akan dilanjutkan pada langkah selanjutnya.
4. Menghilangkan *derivational prefix* yakni “be-”, “di-”, “me-”, “-pe”, “-ter” dan lain sebagainya. Apabila kata dasar ditemukan pada basis data maka proses algoritma akan dihentikan, namun apabila ditemukan maka akan dilakukan *recoding*. Proses ini akan dihentikan apabila telah memenuhi beberapa persyaratan berikut :
    a. Terdapat gabungan antara awalan serta akhiran yang tak di ijinkan
    b. Awalan yang di temukan sama dengan awalan yang telah di hilangkan sebelumnya
    c. Tiga awalan telah dihapus

5. Apabila semua tahap telah di jalankan namun kata dasar belum ditemukan dalam kamus, maka algoritma akan melakukan pengembalian kata yang asli sebelum dilakukan *stemming*.

2.5 Pembobotan

Pembobotan atau *term weighting* ialah proses untuk mendapatkan nilai dari *term* yang sebelumnya telah diekstrak (Puspitasari, 2017).

2.5.1 *Term Frequency* (TF) dan Pembobotan TF (Wtf)

Term Frequency (TF) ialah frekuensi kemunculan *term* (kata) dalam suatu dokumen, frekuensi untuk setiap *term* dapat bervariasi oleh karena itu frekuensi kemunculan *term* menjadi atribut penting untuk membedakan dokumen satu sama lain (Xia dan Chai, 2011) sedangkan Wtf ialah suatu proses untuk melakukan perhitungan bobot untuk setiap *term* (kata). Adapun untuk menentukan nilai TF dan Wtf ditunjukan pada persamaan (2.1) (Manning, Raghavan dan Schutze, 2009).

$$Wf_{t,d} = \begin{cases} 1 + log\ tf_{t,d} & tf_{t,d} \geq 1 \\ 0 & otherwise \end{cases} \quad (2.1)$$

Keterangan :

- $W_{tf_{t,d}}$: Hasil dari pembobotan $tf_{t,d}$
- $tf_{t,d}$: Frekuensi kemunculan *t* pada dokumen *d*

2.5.2 *Document Frequency* (DF_t) dan *Inverse Document Frequency* (IDF_t)

Document Frequency (DFt) ialah jumlah dokumen yang memiliki *term* (kata) *t*, dan *Inverse Document Frequency* ialah jumlah dari dokumen yang memiliki *term* (kata) *t* yang dicari dalam kumpulan dokumen yang ada, persamaan (2.2) merupakan persamaan untuk melakukan perhitungan IDF yang diusulkan oleh Jones (1972) (Manning, Raghavan dan Schutze, 2009).

$$idf_t = log \frac{N}{df_t} \quad (2.2)$$

Keterangan:

- idf_t : Hasil dari invers df_t
- df_t : Jumlah dokumen yang memiliki *t*
- N : Banyak dokumen yang ada

2.5.3 Pembobotan TF-IDF ($W_{t,d}$)

Pembobotan TF-IDF ($W_{t,d}$) ialah proses untuk melakukan penggabungan bobot pada tiap *term* dalam setiap dokumen. Untuk menghitung pembobotan TF-IDF dapat dilakukan dengan melakukan perkalian TF dan IDF_t, untuk menghitung nilai $W_{t,d}$ dapat menggunakan persamaan (2.3) (Manning, Raghavan dan Schutze, 2009).

$$W_{t,d} = W_{tf_{t,d}} * idf_t \quad (2.3)$$

2.5.4 Normalisasi

Normalisasi dilakukan untuk mempermudah dalam menghitung nilai *cosine similarity*, adapun persamaan yang dapat digunakan untuk melakukan normalisasi dapat dilihat pada persamaan (2.4) (Nathania dkk, 2017).

$$w_{t,d} = \frac{w_{t,d}}{\sqrt{\sum_{t=1}^{n} w_{t,d}{}^2}} \quad (2.4)$$

2.5.5 Cosine Similarity

Cosine Similarity ialah metode untuk melakukan perhitungan tingkat kemiripan antar objek. Adapun persamaan untuk melakukan *perhitungan Cosine Similarity* dapat dilihat pada persamaan (2.5) dimana sebelumnya telah dilakukan normalisasi dengan menggunakan persamaan (2.4) (Nathania dkk, 2017).

$$CosSim\left(d_j, q\right) = \overrightarrow{d_j}.\vec{q} = \sum_{i=0}^{t}\left(w_{ij}.w_{iq}\right) \quad (2.5)$$

Keterangan:

- d_j = dokumen latih
- q = dokumen uji
- w_{ij} = nilai hasil pembobotan TF-IDF dokumen latih
- w_{iq} = nilai hasil pembobotan TF-IDF dokumen uji

2.6 K-Nearest Neighbor (K-NN)

K-NN merupakan metode yang sering digunakan untuk melakukan klasifikasi, proses klasifikasi dilakukan dengan menghitung kemiripan antar data uji dengan data latih dan melakukan pertimbangan pada nilai kemiripan tertinggi sejumlah *k* (Zheng et al.,2015).

Algoritma K-NN, pada tahap pembelajarannya menyimpan vektor-vektor fiktur serta klasifikasi dari data pembelajran, dalam tahap klasifikasi fitur-fitur yang mirip akan dihitung sebagai data uji (data dengan klasifikasi yang belum diketahui). Jarak dari vektor baru pada setiap vektor data pembelajaran akan dihitung, serta nilai *k* yang paling dekat akan diambil. Pada algoritma K-NN nilai *k* yang terbaik bergantung kepada data, biasanya nilai *k* yang tinggi dapat memberi pengurangan terhadap efek *noise* dalam klasifikasi, namun membuat *boundary* antar tiap klasifikasi akan lebih kabur. Nilai *k* yang tepat dapat ditentukan dari optimasi parameter, seperti dengan menggunakan *cross-validation*. Pada algoritma K-NN klasifikasi dilakukan dengan memprediksi data pembelajaran yang memiliki jarak paling dekat dengan nilai *k* (Prayoga, Pinandito dan Perdana, 2017).

2.7 Improved K-NN

Metode Improved K-NN adalah metode yang dikembangkan dari metode K-NN, dimana perbedaannya terdapat pada penentuan nilai *k*. Pada K-NN menentukan nilai *k* harus sesuai dan tepat agar bisa mendapatkan nilai akurasi yang tinggi dalam melakukan klasifikasi dokumen.

Pada Improved K-NN dilakukan modifikasi (perubahan) dalam menentukan nilai *k*, yang mana pada Improved K-NN setiap kategorinya memiliki nilai *k* yang berbeda sesuai dengan besar atau kecilnya dokumen latih yang dimiliki oleh setiap kategori, sehingga saat nilai *k* semakin tinggi tidak akan mempengaruhi pada kategori dengan jumlah dokumen latih yang besar (Herdiawan, 2015).

Setelah menghitung nilai *cosine similarity* maka hasil perhitungannya akan diurutkan secara menurun untuk setiap kategori. Setelah itu dilakukan penentuan nilai *k*, selanjutnya akan dilakukan perhitungan untuk mendapatkan nilai *k* baru (*n*), menentukan nilai *k* baru (*n*) dapat dihitung menggunakan persaman (2.6) (Herdiawan, 2015).

$$n = \frac{k * N(c_m)}{Maks[N(c_m)|j = 1..N_c]} \quad (2.6)$$

Keterangan:

- *n* = nilai *k* baru
- *k* = nilai *k* yang ditetapkan
- $N(c_m)$ = banyak dokumen latih kategori m
- $Maks[N(c_m)|j = 1..N_c]$ = banyak dokumen latih terbanyak pada semua kategori

Selanjutnya dilakukan perhitungan peluang dari dokumen uji *X* termasuk dengan dokumen latih d_j sebanyak nilai *n* tetangga untuk setiap kategori pada dokumen *X* pada dokumen latih d_j sebanyak nilai *n* tetangga untuk *training set*. Persamaan (2.7) dapat digunakan untuk menghitung peluang dari dokumen uji *X* pada kategori *m* (Baoli, Shiwen dan Qin, 2003).

$$p(x, c_m) = argMaks_m = \frac{\sum d_j \in top\_n\_kNN\ (c_m) sim(x, d_j) y(d_j, c_m)}{\sum d_j \in top\_n\_kNN\ (c_m) sim(x, d_j)} \quad (2.7)$$

Keterangan:

- $p(x, c_m)$: probabilitas dokumen X anggota c_m
- $sim(x, d_j)$: kemiripan antara dokumen X dengan dokumen latih d_j
- top_n_kNN : nilai n terbaik tetangga
- $y(d_j, c_m)$: fungsi atribut yang memenuhi dari salah satu kategori, apabila dokumen latih $d_j$masuk dalam kategori c_m maka akan bernilai 1, dan sebaliknya jika tidak maka akan bernilai 0.

Setelah menghitung peluang pada dokumen uji *X* pada kategori *m* maka akan dilakukan perbandingan dari hasil peluang pada setiap kategori, nilai peluang terbesar akan menjadi acuan untuk hasil kategori data uji.

2.8 Evaluasi

Evaluasi dilakukan untuk memeriksa keakuratan dari hasil klasifikasi yang dilakukan. Pada pengujian akan diperiksa hasil dari klasifikasi seberapa tingkat akurasi tiap kategori.

2.9 Confusion Matrix

Confusion matrix merupakan salah satu alat yang bisa digunakan untuk melakukan pengujian atau menganalisis hasil klasifikasi. Evaluasi dilakukan dengan menggunakan tabel *confusion matrix* untuk melakukan perbandingan antara kategori aktual dan kategori prediksi (Manning, Raghava dan Schutze, 2009). Tabel 2.1 menunjukan tabel *confusion matrix* (Ting K.M, 2017).

Tabel 2.1 *Confusion Matrix*

Actual Class	*Assigned Class*	
	Positive	*Negative*
Positive	TP	FP
Negative	FN	TN

Dari Tabel 2.1 dapat dilihat bahwa terdapat *actual class* dimana merupakan kategori aktual dan *assigned class* yang merupakan prediksi kategori . pada tabel terdapat nilai TP, FP, FN dan TN. Dimana TP, FP, FN, dan TN memiliki arti sebagai berikut (Puspitasari, 2017) :

- TP (*True Positive*) menunjukan banyak data uji yang masuk dalam kategori x, dan data tersebut benar termasuk kategori x.
- FP (*False Positive*) menunjukan banyak data uji yang tidak masuk kategori x, dan data tersebut seharusnya masuk kategori x.
- FN (*False Negative*) menunjukan banyak data uji masuk dalam kategori x, dan data tersebut seharusnya bukan kategori x.
- TN (*True Negative*) menunjukan banyak data uji tidak masuk dalam kategori x, dan data tersebut memang bukan kategori x.

2.10 *Precision, Recall*, dan F1-*Measure*

2.10.1 Precision

Precision merupakan nilai keakuratan dari hasil pengklasifikasian seluruh dokumen oleh sistem, nilai *precision* dapat dihitung dengan menggunakan persamaan (2.8) (Puspitasari, 2017)

$$Precision = TP/(TP + FP) \quad (2.8)$$

2.10.2 Recall

Recall merupakan tingkat kesuksesan dari sistem untuk mengenali sebuah kategori, nilai dari *recall* dapat dihitung menggunakan persamaan (2.9) (Puspitasari, 2017)

$$Recall = TP/(TP + FN) \quad (2.9)$$

2.10.3 F1-*Measure*

F1-*Measure* ialah sebuah gambara dalam pengaruh *relative* antara *precision* dan *recall*. Dalam menentukan F1-*Measure* dapat menggunakan persamaan (2.10) (Puspitasari, 2017)

$$F - Measure = (2 * P) * R/(P + R) \quad (2.10)$$

BAB 3 METODOLOGI PENELITIAN

Dalam bab ini dijelaskan metode, teknik serta proses yang digunakan dalam penelitian Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN.

3.1 Tipe Penelitian

Pada penelitian Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN memanfaatkan penelitian dengan tipe non-implementatif yang merupakan tipe penelitian dimana melakukan analisis yang kemudian akan di bahas sehingga mendapatkan hasil yang berupa suatu analisis ilmiah. Hasil dari tipe penelitian ini dapat dimanfaatkan untuk melakukan eksperimen, studi kasus, *survey*, dan lain sebagainya.

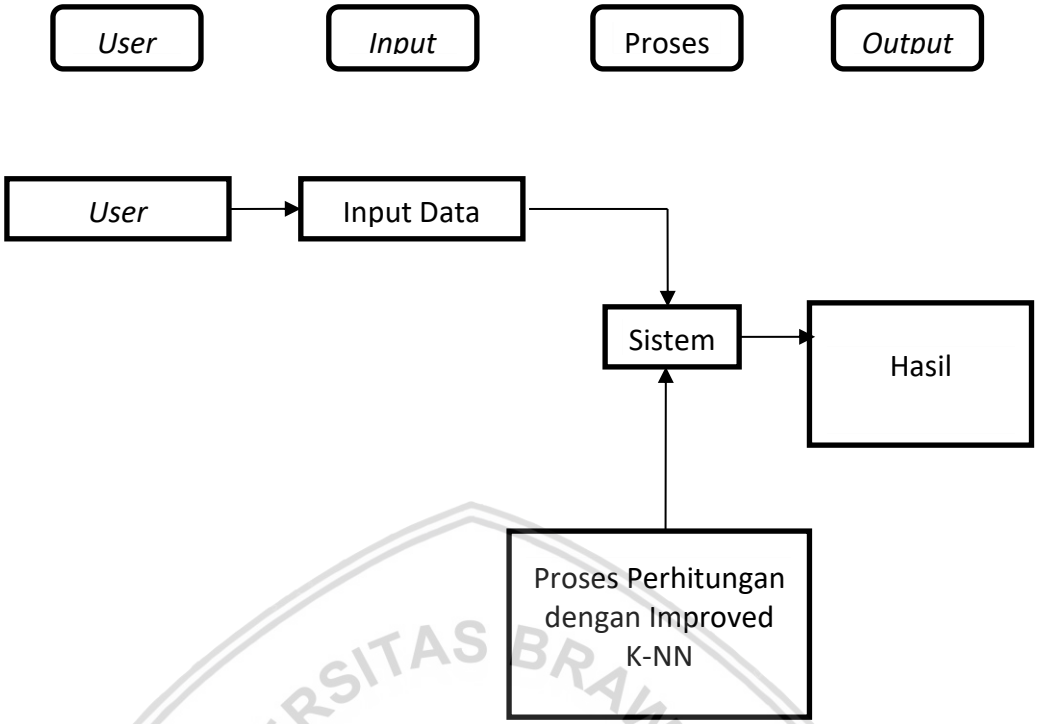
Jenis kegiatan yang dilakukan pada penelitian Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN ialah melakukan pemanfaatan penelitian dengan tipe *analytical* atau *explanatory* yakni penelitian dengan memprioritaskan pada tahap-tahap eksploitasi atau penggalian informasi yang memiliki tujuan untuk melakukan identifikasi bagian penting dari objek penelitian untuk menjadi basis dalam melakukan pengambilan keputusan.

3.2 Strategi Penelitian

Strategi penelitian yang digunakan dalam penelitian Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN ialah strategi penelitian yang berupa eksperimen, dimana penelitian melakukan fokus pada penggunaan satu ataupun lebih variabel dengan cara tertentu sehingga memberikan pengaruh pada variable yang terikat lainnya yang dapat diukur, guna melakukan pengujian pada hipotesis yang memiliki hubungan dengan sebab dan akibat. Adapun algoritma yang digunakan pada penelitian ini ialah algoritma Improved K-NN yang merupakan algoritma pengembangan dari K-NN.

3.3 Rancangan Penelitian

Rancangan sistem digunakan dalam memberikan gambaran tentang cara kerja sistem secara lengkap dan jelas. Cara kerja sistem dimulai dengan memasukan data yang berupa sinopsis film, kemudian data yang dimasukan tersebut akan melewati beberapa tahap, yakni data akan melewati proses pemrosesan data yang dimulai dengan melakukan *preprocessing*, melakukan pembobotan *term* (kata), hingga melakukan proses pengklasifikasian pada data yang dimasukan dengan menggunakan metode Improved K-NN. Setelah itu maka akan didapatkan hasil atau *output* yang dihasilkan oleh sistem yang berupa kategori yang sesuai untuk data yang diujikan. Pada Gambar 3.1 menunjukan perancangan arsitektur pada sistem.



Gambar 3.1 Arsitektur Perancangan Sistem

Pengguna akan memasukan data berupa sinopsis film, dimana data tersebut akan di proses oleh sistem mulai dari *preprocessing* hingga pengklasifikasian menggunakan metode Improved K-NN. Setelah melakukan pengklasifikasian maka didapatkan hasil berupa kategori-kategori genre film yang sesuai.

3.3.1 Partisipan Penelitian

Pada penelitian ini memiliki beberapa pihak yang terlibat yakni, para penulis yang menuliskan sinopsis film yang disediakan pada beberapa situs online yakni sinopsisfilem21.com, posfilm.com, filmbioskop.co.id, pusatsinopsis.com, filmbioskop.net, hype.idntimes.com, filmbor.com, sinopsisfilm.co.id, sinopsisdanreviewfilm.blogspot.com, dan industry.co.id.

3.3.2 Lokasi Penelitian

Penelitian dilakukan dengan mengakses beberapa situs online yang menyediakan sinopsis film. Namun proses-proses pembelajaran dilakukan di Fakultas Ilmu Komputer (FILKOM) Universitas Brawijaya yang berlokasi di Malang, Jawa Timur.

3.3.3 Teknik Pengumpulan Data

Teknik pengumpulan data melakukan pemanfaatan teknik sekunder, yakni menggunakan *dataset* berupa data yang sudah tersedia serta merupakan dokumen yang telah ditulis pada laporan orang lain. Data di dapatkan dari beberapa situs online yakni sinopsisfilem21.com, posfilm.com, filmbioskop.co.id,

pusatsinopsis.com, filmbioskop.net, hype.idntimes.com, filmbor.com, sinopsisfilm.co.id, sinopsisdanreviewfilm.blogspot.com, dan industry.co.id. Data berupa sinopsis film yang ada pada situs online tersebut, nantinya sinopsis tersebut akan diproses dengan memanfaatkan metode Improved K-NN untuk dilakukan klasifikasi sehingga mendapatkan hasil klasifikasi yang akurat.

3.3.4 Teknik Pengujian

Pengujian dilakukan untuk mengetahui tingkat keakuratan dari aplikasi yang dibangun, serta mengetahui apakah aplikasi yang dibangun sesuai dengan tujuan yang diinginkan dalam penelitian yang dilakukan. Adapun pengujian yang dilakukan ialah dengan mengevaluasi hasil kategori dari klasifikasi yang didapatkan dengan menggunakan beberapa data yang digunakan untuk menguji aplikasi. Sehingga dapat menganalisa efektifitas metode Improved K-NN yang digunakan untuk menghasilkan kategori-kategori, serta untuk dapat mengetahui apakah hasil klasifikasi sesuai dengan kategori-kategori yang dihasilkan.

3.3.5 Peralatan Pendukung

Berikut adalah spesifikasi kebutuhan sistem yang dibutuhkan, meliputi perangkat keras maupun lunak. Adapun kebutuhan sistem yang dibutuhkan diantaranya ialah :

1. Perangkat keras (*hardware*) :
   - PC
   - RAM : 2 GB/ 4 GB/ 8 GB
   - ROM : 500 GB
2. Perangkat lunak (*software*) :
   - Editor : Notepad++
   - Database : MySQL
   - Server : Xampp
   - Web Client : Chrome

3.4 Penarikan Kesimpulan dan Saran

Kesimpulan akan dilakukan setelah semua tahapan perancangan, implementasi dan pengujian metode yang diterapkan telah selesai dilakukan, kesimpulan didapatkan dari hasil pengujian dan analisis metode yang diterapkan.

3.5 Jadwal Penelitian

Penelitian dilakukan sejak bulan Februari hingga Agustus tahun 2018, adapun jadwal dari penelitian ini dapat dilihat pada Tabel 3.1

Tabel 3.1 Jadwal Penelitian

No	Uraian	Minggu ke-																											
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
		Februari				Maret				April				Mei				Juni				Juli				Agustus			
1	Konsultasi dan Penyusunan Laporan																												
2	Penyerahan Proposal																												
3	Perbaikan/ Revisi Proposal Penelitian																												
4	Pengumpulan Data																												
5	Penyusunan Laporan Penelitian																												
6	Bimbingan dan Konsultasi Hasil Penelitian																												
7	Seminar Hasil Penelitian																												
8	Perbaikan Hasil Penelitian																												
9	Sidang Skripsi																												
10	Perbaikan Hasil Sidang Penelitian																												

BAB 4 PERANCANGAN DAN IMPLEMENTASI

Pada bagian ini dilakukan pembahasan mengenai analisis serta perancangan sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN.

4.1 Deskripsi Masalah

Film merupakan salah satu media untuk berkomunikasi yang memiliki sifat audio visual dimana tersirat pesan yang ingin disampaikan oleh pencipta film. Pesan pada film dapat berbeda tergantung dari genre film, umumnya pesan yang disampaikan oleh film berupa pesan tentang pendidikan, informasi, serta untuk hiburan. Secara umum film dapat digolongkan menjadi dua golongan yaitu film fiksi dan non fiksi. Film fiksi diciptakan berdasarkan kisah yang tidak nyata atau dikarang, sedangkan film non fiksi diciptakan berdasarkan kisah nyata. Dari dua golongan film tersebut dilahirkanlah beberapa genre film yang beragam yakni romantis, horor, *thriller*, komedi, fantasi dan lain sebagainya. Dari banyaknya genre yang disediakan oleh film, tak sedikit penikmat film yang masih bingung akan perbedaan dari genre-genre tersebut.

Tidak sedikit penikmat film yang masih bingung membedakan atau menentukan genre film yang sesuai dengan yang diinginkan, serta agar pesan pada film dapat ditujukan dan disampaikan dengan tepat maka melakukan pengelompokan film atau klasifikasi pada sinopsis film dirasa menjadi solusi yang tepat untuk masalah tersebut. Melakukan klasifikasi pada sinopsis film akan membantu dalam mengelompokan film dengan genre yang sesuai. Penelitian ini menggunakan metode Improved K-NN (K-Nearest Neighbor).

Metode K-NN ialah metode untuk melakukan pengelompokan objek sesuai dengan jarak yang paling dekat dari objek dengan masing-masing kategori (Sreemathy dan Balamurugan, 2012). Namun dalam penerapannya metode K-NN memiliki kekurangan yaitu ketika melakukan penentuan kelas dari data kandidat hasil yang didapat masih kurang tepat, maka dengan menggunakan metode Improved K-NN dapat menjadi solusi yang tepat untuk mengatasi masalah tersebut (Megantara et all, 2010). Perbedaan antara metode K-NN dan Improved K-NN terdapat pada penentuan nilai *k*, pada K-NN nilai *k* yang ditentukan pada tiap kategori ialah memiliki nilai yang sama, sedangkan pada Improved K-NN digunakan nilai *k* yang berbeda pada tiap kategori yang sesuai dengan banyaknya data latih (Puspitasari et all, 2017). Sehingga nilai akurasi yang didapatkan akan lebih tinggi dan maksimal. Metode ini dirasa tepat untuk melakukan klasifikasi sehingga dapat menghasilkan kelas-kelas yang sesuai.

4.2 Deskripsi Umum Sistem

Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved-KNN ialah sistem yang dikembangkan untuk melakukan pengklasifikasian dokumen berupa sinopsis film yang termasuk dalam beberapa kategori yaitu romantis,

horor, aksi, *thriller*, dan keluarga, pengguna sistem dapat memberikan masukan berupa dokumen yang beris sinopsis film. Hasil klasifikasi yang didapatkan akan dipengaruhi oleh banyaknya data latih yang digunakan yang mana telah melewati beberapa proses yaitu *preprocessing*, pembobotan *term*, normalisasi serta hasil dari pembobotan *term*. Penentuan kelas atau kategori dari data uji di di *input*kan akan mengacu pada data latih yang sebelumnya telah diproses terlebih dahulu serta proses klasifikasi akan dilakukan menggunakan metode Improved K-NN.

4.3 Manualisasi Perhitungan Data

Manualisasi perhitungan data ini akan menampilkan perhitungan dalam melakukan proses klasifikasi dari dokumen data latih dan data uji.

4.3.1 Preprocessing

Pada tahap *preprocessing* dilakukan beberapa tahapan yaitu *cleansing, case folding*, tokenisasi, *filtering*, dan *stemming*. Hal tersebut dilakukan agar data terstruktur sehingga mempermudah dalam proses perhitungan. Tabel 4.1 menunjukan data latih yang digunakan, dan pada Tabel 4.2 menunjukan data uji yang digunakan.

Tabel 4.1 Data latih

Judul	Sinopsis	Kategori
Galih & Ratna	film yang bercerita tentang kisah dari sepasang kekasih yang ikonik dari sebuah novel dengan judul gita cinta di sma karya dari eddy d iskandar film ini mengisahkan perjalanan cinta dari dua remaja di penghujung sma antara galih refal hadi dan ratna sheryl sheinafia	Romantis
Beauty and The Beast	film yang diadaptasi dari sebuah dongeng yang menceritakan tentang seorang pangeran monster dan seorang wanita muda yang saling jatuh cinta	Romantis
Strong	film yang bercerita berdasarkan kisah nyata yang menceritakan mengenai sekelompok pasukan elit khusus dan para agen cia mereka secara diamdiam menyerang afghanistan pasca tragedy dengan menunggang kuda dan membantu para pejuang afghanistan mereka merebut kota mazarisharif dan menjatuhkan taliban	Aksi
Kingsman: The Golden Circle	film yang bercerita tentang agen rahasia dari kingsman setelah markas besar kingsman dihancurkan oleh poppy	Aksi

	julianne moore yang merupkan seorang penjahat terkenal dan anggota baru dari kelompok rahasia yang dinamakan the golden circle membuat gary eggsy unwin taron egerton merlin mark strong dan roxy sophie cookson yang merupakan anggota agen rahasia kingsman harus pergi ke amerika serikat untuk bergabung dengan rekanrekan kingsman lainnya di sana	
Danur : Maddah	film yang bercerita tentang kisah dari kelanjutan kisah mengenai raisa prilly latuconsina bersama dengan para sahabat hantunya yaitu : william peter hans janshen dan hendrik kali ini akan hadir atau muncul dua anggota baru yaitu norma dan marianne namun peter bersama dengan temantemannya merasa kurang nyaman karena kehadiran marianne yang memiliki sifat egois dan keras	Horor
V.I.P	film yang bercerita tentang kisah dari seorang lelaki dan juga anak lelaki dari pejabat tinggi korea utara yang bernama kwang il lee jong suk yang melakukan pembunuhan pada beberapa bagian di negaranegara termasuk negara tetangga korea selatan karena itu dia akhirnya diburu oleh tim penyidik korsel korut dan pihak interpol untuk mempertanggung jawabkan perbuataannya	*Thriller*
Wonder	film wonder bercerita tentang seorang anak lakilaki bernama august pullman jacob trembley yang biasa dipanggil auggie terlahir dengan kelainan bentuk wajah yang sangat langka yang dikenal sebagai mandibulofacial dysostosis yang kemungkinan merupakan sindrom treacher collins hal itu membuat auggie minder dan menghindar untuk pergi ke sekolah umum selama operasi wajah auggie belajar di rumah dengan metode homeschooling oleh ibunya isabel julia roberts	Keluarga
Ayah Menyayangi Tanpa Akhir	film yang diangkat dari novel yang menjadi best seller karya dari kirana kejora dengan judul yang sama juga	Keluarga

	menceritakan awal kisah cinta antara sepasang kekasih namun memiliki perbedaan asal keturunan dengan berbagai langkah perjuangan akhirnya juna dan keisya bisa menikah dan memiliki seorang buah hati bernama mada nauval azhar kehadiran sang buah hati memberikan bahagia dan duka karena juna kehilangan keisya saat melahirkan mada dengan ini juna pun menjalankan dua peran sebagai seorang ayah sekaligus ibu bagi mada saat dewasa mada divonis menderita kanker otak juna dengan cinta dan kasih sayangnya yang begitu besar akan mada berjuang akan kesembuhan mada namun pada akhirnya mada pun meninggal dan pesan mada terhadap sang ayah sebelum meninggal agar juna tetap untuk bisa terus melanjutkan hidupnya	

Tabel 4.2 Data uji

Judul	Sinopsis
A Family Man	A Family Man berkisah tentang Dane Jensen (Gerard Butler) seorang headhunter yang bekerja pada agency Blackrock Recruting akhirnya mencapai tujuan lamanya untuk dapat mengambil alih perusahaan tersebut. Ia berhasil mengalahkan pesaingnya yang ambisius Lynn Vogel (Alison Brie). Namun setelah keinginannya tercapai, ia dihadapkan dengan kenyataan anak laki-lakinya yang berusia 10 tahun, Ryan (Max Jenkins) didiagnosis menderita kanker.Dane mengalami kegamangan, prioritas profesionalnya di tempat kerja dan prioritas pribadi di rumah mulai berbenturan.

4.3.1.1 Cleansing

Tahapan *cleansing* ialah proses yang dilakukan untuk membersihkan *text* pada dokumen dari karakter yang tidak diperlukan. Tabel 4.3 menunjukan hasil setelah melakukan *cleansing* terhadap data latih dan Tabel 4.4 menunjukan hasil setelah melakukan *cleansing* pada data uji.

Tabel 4.3 Hasil *cleansing* data latih

Judul	Sinopsis	Kategori
Galih & Ratna	film yang bercerita tentang kisah dari sepasang kekasih yang ikonik dari	Romantis

	sebuah novel dengan judul gita cinta di sma karya dari eddy d iskandar film ini mengisahkan perjalanan cinta dari dua remaja di penghujung sma antara galih refal hadi dan ratna sheryl sheinafia	
Beauty and The Beast	film yang diadaptasi dari sebuah dongeng yang menceritakan tentang seorang pangeran monster dan seorang wanita muda yang saling jatuh cinta	Romantis
Strong	film yang bercerita berdasarkan kisah nyata yang menceritakan mengenai sekelompok pasukan elit khusus dan para agen cia mereka secara diamdiam menyerang afghanistan pasca tragedy dengan menunggang kuda dan membantu para pejuang afghanistan mereka merebut kota mazarisharif dan menjatuhkan taliban	Aksi
Kingsman: The Golden Circle	film yang bercerita tentang agen rahasia dari kingsman setelah markas besar kingsman dihancurkan oleh poppy julianne moore yang merupkan seorang penjahat terkenal dan anggota baru dari kelompok rahasia yang dinamakan the golden circle membuat gary eggsy unwin taron egerton merlin mark strong dan roxy sophie cookson yang merupakan anggota agen rahasia kingsman harus pergi ke amerika serikat untuk bergabung dengan rekanrekan kingsman lainnya di sana	Aksi
Danur : Maddah	film yang bercerita tentang kisah dari kelanjutan kisah mengenai raisa prilly latuconsina bersama dengan para sahabat hantunya yaitu william peter hans janshen dan hendrik kali ini akan hadir atau muncul dua anggota baru yaitu norma dan marianne namun peter bersama dengan temantemannya merasa kurang nyaman karena kehadiran marianne yang memiliki sifat egois dan keras	Horor
V.I.P	film yang bercerita tentang kisah dari seorang lelaki dan juga anak lelaki dari pejabat tinggi korea utara yang bernama kwang il lee jong suk yang melakukan pembunuhan pada beberapa	*Thriller*

	bagian di negaranegara termasuk negara tetangga korea selatan karena itu dia akhirnya diburu oleh tim penyidik korsel korut dan pihak interpol untuk mempertanggung jawabkan perbuataannya	
Wonder	film wonder bercerita tentang seorang anak lakilaki bernama august pullman jacob trembley yang biasa dipanggil auggie terlahir dengan kelainan bentuk wajah yang sangat langka yang dikenal sebagai mandibulofacial dysostosis yang kemungkinan merupakan sindrom treacher collins hal itu membuat auggie minder dan menghindar untuk pergi ke sekolah umum selama operasi wajah auggie belajar di rumah dengan metode homeschooling oleh ibunya isabel julia roberts	Keluarga
Ayah Menyayangi Tanpa Akhir	film yang diangkat dari novel yang menjadi best seller karya dari kirana kejora dengan judul yang sama juga menceritakan awal kisah cinta antara sepasang kekasih namun memiliki perbedaan asal keturunan dengan berbagai langkah perjuangan akhirnya juna dan keisya bisa menikah dan memiliki seorang buah hati bernama mada nauval azhar kehadiran sang buah hati memberikan bahagia dan duka karena juna kehilangan keisya saat melahirkan mada dengan ini juna pun menjalankan dua peran sebagai seorang ayah sekaligus ibu bagi mada saat dewasa mada divonis menderita kanker otak juna dengan cinta dan kasih sayangnya yang begitu besar akan mada berjuang akan kesembuhan mada namun pada akhirnya mada pun meninggal dan pesan mada terhadap sang ayah sebelum meninggal agar juna tetap untuk bisa terus melanjutkan hidupnya	Keluarga

Tabel 4.4 Hasil *cleansing* data uji

Judul	Sinopsis

A Family Man	A Family Man berkisah tentang Dane Jensen Gerard Butler seorang headhunter yang bekerja pada agency Blackrock Recruting akhirnya mencapai tujuan lamanya untuk dapat mengambil alih perusahaan tersebut Ia berhasil mengalahkan pesaingnya yang ambisius Lynn Vogel Alison Brie Namun setelah keinginannya tercapai ia dihadapkan dengan kenyataan anak laki-lakinya yang berusia tahun Ryan Max Jenkins didiagnosis menderita kanker Dane mengalami kegamangan prioritas profesionalnya di tempat kerja dan prioritas pribadi di rumah mulai berbenturan

4.3.1.2 Case Folding

Tahapan *case folding* ialah proses yang dilakukan untuk mengubah semua huruf pada dokumen menjadi huruf kecil. Tabel 4.5 menunjukan hasil setelah melakukan *case folding* terhadap data latih dan Tabel 4.6 menunjukan hasil setelah melakukan *case folding* pada data uji.

Tabel 4.5 Hasil *case folding* data latih

Judul	Sinopsis	Kategori
Galih & Ratna	film yang bercerita tentang kisah dari sepasang kekasih yang ikonik dari sebuah novel dengan judul gita cinta di sma karya dari eddy d iskandar film ini mengisahkan perjalanan cinta dari dua remaja di penghujung sma antara galih refal hadi dan ratna sheryl sheinafia	Romantis
Beauty and The Beast	film yang diadaptasi dari sebuah dongeng yang menceritakan tentang seorang pangeran monster dan seorang wanita muda yang saling jatuh cinta	Romantis
Strong	film yang bercerita berdasarkan kisah nyata yang menceritakan mengenai sekelompok pasukan elit khusus dan para agen cia mereka secara diamdiam menyerang afghanistan pasca tragedy dengan menunggang kuda dan membantu para pejuang afghanistan mereka merebut kota mazarisharif dan menjatuhkan taliban	Aksi
Kingsman: The Golden Circle	film yang bercerita tentang agen rahasia dari kingsman setelah markas besar kingsman dihancurkan oleh poppy julianne moore yang merupkan seorang penjahat terkenal dan anggota baru dari kelompok rahasia yang dinamakan the	Aksi

	golden circle membuat gary eggsy unwin taron egerton merlin mark strong dan roxy sophie cookson yang merupakan anggota agen rahasia kingsman harus pergi ke amerika serikat untuk bergabung dengan rekanrekan kingsman lainnya di sana	
Danur : Maddah	film yang bercerita tentang kisah dari kelanjutan kisah mengenai raisa prilly latuconsina bersama dengan para sahabat hantunya yaitu william peter hans janshen dan hendrik kali ini akan hadir atau muncul dua anggota baru yaitu norma dan marianne namun peter bersama dengan teman temannya merasa kurang nyaman karena kehadiran marianne yang memiliki sifat egois dan keras	Horor
V.I.P	film yang bercerita tentang kisah dari seorang lelaki dan juga anak lelaki dari pejabat tinggi korea utara yang bernama kwang il lee jong suk yang melakukan pembunuhan pada beberapa bagian di negaranegara termasuk negara tetangga korea selatan karena itu dia akhirnya diburu oleh tim penyidik korsel korut dan pihak interpol untuk mempertanggung jawabkan perbuataannya	*Thriller*
Wonder	film wonder bercerita tentang seorang anak lakilaki bernama august pullman jacob trembley yang biasa dipanggil auggie terlahir dengan kelainan bentuk wajah yang sangat langka yang dikenal sebagai mandibulofacial dysostosis yang kemungkinan merupakan sindrom treacher collins hal itu membuat auggie minder dan menghindar untuk pergi ke sekolah umum selama operasi wajah auggie belajar di rumah dengan metode homeschooling oleh ibunya isabel julia roberts	Keluarga
Ayah Menyayangi Tanpa Akhir	film yang diangkat dari novel yang menjadi best seller karya dari kirana kejora dengan judul yang sama juga menceritakan awal kisah cinta antara sepasang kekasih namun memiliki perbedaan asal keturunan dengan	Keluarga

	berbagai langkah perjuangan akhirnya juna dan keisya bisa menikah dan memiliki seorang buah hati bernama mada nauval azhar kehadiran sang buah hati memberikan bahagia dan duka karena juna kehilangan keisya saat melahirkan mada dengan ini juna pun menjalankan dua peran sebagai seorang ayah sekaligus ibu bagi mada saat dewasa mada divonis menderita kanker otak juna dengan cinta dan kasih sayangnya yang begitu besar akan mada berjuang akan kesembuhan mada namun pada akhirnya mada pun meninggal dan pesan mada terhadap sang ayah sebelum meninggal agar juna tetap untuk bisa terus melanjutkan hidupnya	

Tabel 4.6 Hasil *case folding* data uji

Judul	Sinopsis
A Family Man	berkisah tentang dane jensen gerard butler seorang headhunter yang bekerja pada agency blackrock recruting akhirnya mencapai tujuan lamanya untuk dapat mengambil alih perusahaan tersebut ia berhasil mengalahkan pesaingnya yang ambisius lynn vogel alison brie namun setelah keinginannya tercapai ia dihadapkan dengan kenyataan anak lakilakinya yang berusia tahun ryan max jenkins didiagnosis menderita kanker dane mengalami kegamangan prioritas profesionalnya di tempat kerja dan prioritas pribadi di rumah mulai berbenturan

4.3.1.3 Tokenisasi

Tahapan tokenisasi dilakukan untuk melakukan pemisahan pada setiap kata yang dipisahkan oleh *whitespace* pada dokumen. Tabel 4.7 menunjukan hasil tokenisasi pada data latih dan Tabel 4.8 menunjukan hasil tokenisasi pada data uji.

Tabel 4.7 Hasil tokenisasi data latih

Judul	Sinopsis	Kategori
Galih & Ratna	film yang bercerita tentang kisah dari sepasang kekasih yang ikonik dari sebuah novel dengan judul gita cinta di sma karya dari eddy d iskandar film ini mengisahkan perjalanan cinta dua	Romantis

	remaja di penghujung sma antara galih refal hadi dan ratna sheryl sheinafia	
Beauty and The Beast	film yang diadaptasi dari sebuah dongeng yang menceritakan tentang seorang pangeran monster dan seorang wanita muda yang saling jatuh cinta	Romantis
Strong	film yang bercerita berdasarkan kisah nyata yang menceritakan mengenai sekelompok pasukan elit khusus dan para agen cia mereka secara diam diam menyerang afghanistan pasca tragedy dengan menunggang kuda dan membantu para pejuang afghanistan mereka merebut kota mazarish arif dan menjatuhkan Taliban	Aksi
Kingsman: The Golden Circle	film yang bercerita tentang agenrahasia dari kingsman setelah markas besar kingsman dihancurkan oleh poppy julianne yang merupakan seorang penjahat terkenal dan anggota baru kelompok rahasia yang dinamakan the golden circle membuat gary eggsy unwin taron egerton merlin mark strong dan roxy sophie cookson yang merupakan anggota agen rahasia kingsman harus pergi ke amerika serikat untuk bergabung dengan rekan rekan kingsman lainnya disana	Aksi
Danur : Maddah	film yang bercerita tentang kisah dari kelanjutan kisah mengenai raisa prily latuconsina bersama dengan para sahabat hantunya yaitu william peter hans janshen dan hendrik kali ini akan hadir atau muncul dua anggota baru yaitu norma dan marianne namun peter bersama dengan teman temannya merasa kurangnyaman karena kehadiran marianne yang memiliki sifat egois dan keras	Horor
V.I.P	film yang bercerita tentang kisah dari seorang lelaki dan juga anak lelaki dari pejabat tinggi korea utara yang bernama kwang il lee jong suk yang melakukan pembunuhan pada beberapa bagian di negara negara termasuk negara tetangga korea selatan karena itu dia akhirnya diburu oleh tim	*Thriller*

	penyidik korsel korut dan pihak interpol untuk mempertanggung jawabkan perbuatannya	
Wonder	bercerita tentang seoranganak laki laki bernama august pullman jacob trembley yang biasa dipanggil auggie terlahir dengan kelainan bentuk wajah yang sangat langka yang dikenal sebagai mandibulofacial dysostosis yang kemungkinan merupakan sindrom treacher collins hal itu membuat auggie minder dan menghindar untuk pergi ke sekolah umum selama operasi wajah auggie belajar dirumah dengan metode homeschooling oleh ibunya isabel julia roberts	Keluarga
Ayah Menyayangi Tanpa Akhir	film yang diangkat dari novel yang menjadi best seller karya dari kirana kejora dengan judul yang sama juga menceritaka awal kisah cinta antara sepasang kekasih namun memiliki perbedaan asal keturunan dengan berbagai langkah perjuangan akhirnya juna dan keysa bisa menikah dan memiliki seorang buah hati bernama mada nauval azhar kebadiran sang buah hati memberikan bahagia dan duka karena juna kehilangan keysa saat melahirkan mada dengan ini juna pun menjalankan dua peran sebagai seorang ayah sekaligus ibu bagi mada saat dewasa mada divonis menderita kanker otak juna dengan cinta dan kasih sayangnya yang begitu besar akan mada berjuang akan kesembuhan mada namun pada akhirnya madapun meninggal dan pesan mada terhadap sang ayang sebelum meninggal agar juna tetap untuk bisa terus melanjutkan hidupnya	Keluarga

Tabel 4.8 Hasil tokenisasi data uji

Judul	Sinopsis

A Family Man	berkisah tentang dane jensen gerard butler seorang headhunter yang bekerja pada agency blackrock recruiting akhirnya mencapai tujuan lamanya untuk dapat mengambil alih perusahaan tersebut ia berhasil mengalahkan pesaingnya yang ambisius lynn vogel alison brie namun setelah keinginanya tercapai ia dihadapkan dengan kenyataan anak laki lakinya yang berusia tahun ryan max jenkins di diagnosis menderita kanker dane mengalami kegamangan prioritas profesionalnya di tempat kerja dan prioritas pribadi di rumah mulai berbenturan

4.3.1.4 Filtering

Pada tahap ini dilakukan proses penghapusan kata yang kurang penting yang sesuai dengan stopword. Tabel 4.9 menunjukan hasil *filtering* pada data latih dan Tabel 4.10 menunjukan hasil filtering pada data uji.

Tabel 4.9 Hasil *filtering* data latih

Judul	Sinopsis	Kategori
Galih & Ratna	kisah kekasih ikonik novel gita cinta sma eddy d iskandar mengisahkan perjalanan cinta remaja penghujung sma galih refal hadi ratna sheryl sheinafia	Romantis
Beauty and The Beast	adaptasi dongeng pangeran monster wanita muda jatuh cinta	Romantis
Strong	sekelompok pasukan elit khusus agen cia diam diam menyerang afghanistan pasca tragedy menunggang kuda membantu pejuang afghanistan merebut kota mazarish arif menjatuhkan Taliban	Aksi
Kingsman: The Golden Circle	agen rahasia kingsman markas kingsmn dihancurkan poppy julianne penjahat terkenal anggota baru kelompok rahasia the golden circle membuat gry eggsy unwin taron egerton merlin mark strong roxy sophie cookson anggota agen rahasia kingsman pergi amerika serikat bergabung rekan rekan kingsman	Aksi
Danur : Maddah	mengenai raisa prily latuconsina sahabat hantunya william peter hans janshen hendrik hadir muncul dua anggota baru norma mariane peter bersama teman temannya kurang	Horor

	nyaman kehadiran marianne memiliki sifat egois keras	
V.I.P	lelaki anak lelaki pejabat tinggi korea utara kwang il lee jong suk melakukan pembunuhan bagian negara negara tetangga korea selatan diburu tim penyidik korsel korut pihak interpol mempertanggung jawabkan perbuatannya	*Thriller*
Wonder	anak laki laki august pullman jacob trembley dipanggil auggie terlahir kelainan bentuk wajah langka mandibulofacial dysostosis sindrom treacher collins auggie minder menghindar sekolah umumoperasi wajah auggie belajar dirumah dengan metode homeschooling ibunya isabel julia roberts	Keluarga
Ayah Menyayangi Tanpa Akhir	novel best seller kirana kejora kisah cinta kekasih memiliki perbedaan keturunan langkah perjuangan juna keysa menikah memiliki buah hati mada nauval azhar kehadiran sang buah hati memberikan bahagia duka juna kehilangan keysa melahirkan mada juna menjalankan peran ayah ibu mada dewasa mada divonis menderita kanker otak juna cinta kasih sayangnya besar mada berjuang kesembuhan mada madapun meninggal pesan mada sang ayah meninggal juna melanjutkan hidupnya	Keluarga

Tabel 4.10 Hasil *filtering* data uji

Judul	Sinopsis
A Family Man	berkisah dane jensen gerard butler headhunter bekerja agency blackrock recruiting mencapai tujuan lamanya mengambil alih perusahaan berhasil mengalahkan pesaingnya ambisius lynn vogel alison brie keinginanya tercapai dihadapkan kenyataan anak laki lakinya berusia tahun ryan max jenkins didiagnosis menderita kanker dane mengalami kegamangan prioritas profesionalnya tempat kerja prioritas pribadi rumah berbenturan

4.3.1.5 Stemming

Pada tahap ini dilakukan perubahan pada tiap kata di dokumen menjadi kata dasar. Tabel 4.11 menunjukan hasil *stemming* pada data latih dan Tabel 4.12 menunjukan hasil *stemming* pada data uji.

Tabel 4.11 Hasil *stemming* pada data latih

Judul	Sinopsis	Kategori
Galih & Ratna	kisah kasih ikonik novel gita cinta sma eddy d iskandar kisah jalan cinta remaja ujung sma galih refal hadi ratna sheryl sheinafia	Romantis
Beauty and The Beast	adaptasi dongeng pangeran monster wanita muda jatuh cinta	Romantis
Strong	kelompok pasuk elit khusus agen cia diam diam serang afghanistan pasca tragedy tunggang kuda bantu juang afghanistan rebut kota mazarish arif jatuh Taliban	Aksi
Kingsman: The Golden Circle	agen rahasia kingsman markas kingsmn hancur poppy julianne jahat kenal anggota baru kelompok rahasia the golden circle buat gray eggsy unwin taron egerton merlin mark strong roxy sophie cookson anggota agen rahasia kingsman pergi amerika serikat gabung rekan rekan kingsman	Aksi
Danur : Maddah	kena raisa prily latuconsina sahabat hantu william peter hans janshen hendrik hadir muncul dua anggota baru norma mariane peter sama teman teman kurang nyaman hadir marianne milik sifat egois keras	Horor
V.I.P	laki anak laki jabat tinggi korea utara Kwang il lee jong suk laku bunuh bagian negara negara tetangga korea selatan buru tim sidik korsel korut pihak Interpol tanggung jawab buat	*Thriller*
Wonder	anak laki laki august pullman jacob trembley panggil auggie lahir lain bentuk wajah langka mandibulofacial dysostosis sindrom treacher collins auggie minder hindar sekolah umum operasi wajah auggie belajar rumah metode homeschooling ibu sabel julia roberts	Keluarga

Ayah Menyayangi Tanpa Akhir	novel best seller kirana kejora kisah cinta kasih milik beda turun langkah juang juna keysa nikah milik buah hati mada nauval azhar hadir sang buah hati beri bahagia duka juna hilang keysa lahir mada juna jalan peran ayah ibu mada dewasa mada vonis derita kanker otak juna cinta kasih sayang besar mada juang sembuh mada mada tinggal pesan mada sang ayah tinggal juna lanjut hidup	Keluarga

Tabel 4.12 Hasil *stemming* pada data uji

Judul	Sinopsis
A Family Man	kisah dane jensen gerard butler headhunter kerja agency blackrock recruiting capai tuju lama ambil alih usaha hasil kalah saing ambisius lynn vogel alison brie ingin capai hadap nyata anak laki laki usia tahun ryan max jenkins diagnosis derita kanker dane alami gamang prioritas profesional tempat kerja prioritas pribadi rumah bentur

4.3.2 Pembobotan

Pada tahap ini, TF-IDF dilakukan untuk menghitung nilai pembobotan yang dimana nilai IDF perbedaan dalam kemunculan kata pada dokumen yang ada. Persamaan yang digunakan ditunjukan pada persamaan 2.1 hingga 2.3.

4.3.2.1 Menghitung TF dan IDF

Perhitungan TF dan IDF dimulai dengan melakukan perhitungan pada banyaknya kemunculan *term* (kata) pada dokumen dengan menggunakan persamaan (2.1). Berikut contoh untuk melakukan perhitungan TF pada *term* kisah pada d1.

$$W_{tf_{t,d}} = 1 + log_{10} tf_{t,d} = 1 + log_{10} 1 = 1 + 0 = 1$$

Setelah itu melakukan perhitungan untuk mengetahui banyaknya dokumen *d* yang memiliki *t* (df), kemudian melakukan perhitungan IDF dengan menggunakan persamaan (2.2). Berikut contoh untuk melakukan perhitungan IDF pada *term* kisah pada d1.

$$idf_t = log \frac{8}{5} = \ 0{,}20412$$

Pada Tabel 4.13 menunjukan hasil dari perhitungan TF dan IDF.

Tabel 4.13 Hasil perhitungan TF dan IDF

Term	Tf									df	idf
	d1	d2	d3	d4	d5	d6	d7	d8	Uji		
kisah	1	0	1	0	1,301	1	0	1	1	5	0,20412
kasih	1	0	0	0	0	0	0	1,301	0	2	0,60206
ikonik	1	0	0	0	0	0	0	0	0	1	0,90309
gita	1	0	0	0	0	0	0	0	0	1	0,90309
cinta	1,301	1	0	0	0	0	0	1,301	0	3	0,42597
sma	1,301	0	0	0	0	0	0	0	0	1	0,90309
jalan	1	0	0	0	0	0	0	1	0	2	0,60206
remaja	1	0	0	0	0	0	0	0	0	1	0,90309
ujung	1	0	0	0	0	0	0	0	0	1	0,90309
dongeng	0	1	0	0	0	0	0	0	0	1	0,90309
pangeran	0	1	0	0	0	0	0	0	0	1	0,90309
monster	0	1	0	0	0	0	0	0	0	1	0,90309
wanita	0	1	0	0	0	0	0	0	0	1	0,90309
muda	0	1	0	0	0	0	0	0	0	1	0,90309
jatuh	0	1	1	0	0	0	0	0	0	2	0,60206
cinta	1,301	1	0	0	0	0	0	1,301	0	3	0,42597
kelompok	0	0	1	1	0	0	0	0	0	2	0,60206
pasuk	0	0	1	0	0	0	0	0	0	1	0,90309
elit	0	0	1	0	0	0	0	0	0	1	0,90309
agen	0	0	1	1,301	0	0	0	0	0	2	0,60206
cia	0	0	1	0	0	0	0	0	0	1	0,90309
serang	0	0	1	0	0	0	0	0	0	1	0,90309
afghanistan	0	0	1	0	0	0	0	0	0	1	0,90309
bantu	0	0	1	0	0	0	0	0	0	1	0,90309
juang	0	0	1	0	0	0	0	0	0	1	0,90309
rebut	0	0	1	0	0	0	0	0	0	1	0,90309
sahabat	0	0	0	0	1	0	0	0	0	1	0,90309
hantu	0	0	0	0	1	0	0	0	0	1	0,90309
hadir	0	0	0	0	1,301	0	0	0	0	1	0,90309
anggota	0	0	0	1,301	1	0	0	0	0	2	0,60206
teman	0	0	0	0	1,301	0	0	0	0	1	0,90309
nyaman	0	0	0	0	1	0	0	0	0	1	0,90309
sifat	0	0	0	0	1	0	0	0	0	1	0,90309
egois	0	0	0	0	1	0	0	0	0	1	0,90309
keras	0	0	0	0	1	0	0	0	0	1	0,90309
laki	0	0	0	0	0	1,301	1,301	0	1,301	2	0,60206
anak	0	0	0	0	0	1	1	0	1	2	0,60206
jabat	0	0	0	0	0	1	0	0	0	1	0,90309
korea	0	0	0	0	0	1	0	0	0	1	0,90309

bunuh	0	0	0	0	0	1	0	0	0	1	0,90309
negara	0	0	0	0	0	3	0	0	0	1	0,90309
buru	0	0	0	0	0	1	0	0	0	1	0,90309
tim	0	0	0	0	0	1	0	0	0	1	0,90309
sidik	0	0	0	0	0	1	0	0	0	1	0,90309
interpol	0	0	0	0	0	1	0	0	0	1	0,90309
tanggung	0	0	0	0	0	1	0	0	0	1	0,90309
jawab	0	0	0	0	0	1	0	0	0	1	0,90309
lahir	0	0	0	0	0	0	1	1	0	2	0,60206
bentuk	0	0	0	0	0	0	1	0	0	1	0,90309
wajah	0	0	0	0	0	0	2	0	0	1	0,90309
langka	0	0	0	0	0	0	1	0	0	1	0,90309
minder	0	0	0	0	0	0	1	0	0	1	0,90309
sekolah	0	0	0	0	0	0	1	0	0	1	0,90309
operasi	0	0	0	0	0	0	1	0	0	1	0,90309
ajar	0	0	0	0	0	0	1	0	0	1	0,90309
rumah	0	0	0	0	0	0	1	0	1	1	0,90309
ibu	0	0	0	0	0	0	1,301	1	0	2	0,60206
nikah	0	0	0	0	0	0	0	1	0	1	0,90309
buah	0	0	0	0	0	0	0	2	0	1	0,90309
hati	0	0	0	0	0	0	0	2	0	1	0,90309
bahagia	0	0	0	0	0	0	0	1	0	1	0,90309
duka	0	0	0	0	0	0	0	1	0	1	0,90309
hilang	0	0	0	0	0	0	0	1	0	1	0,90309
peran	0	0	0	0	0	0	0	1	0	1	0,90309
ayah	0	0	0	0	0	0	0	2	0	1	0,90309
dewasa	0	0	0	0	0	0	0	1	0	1	0,90309
vonis	0	0	0	0	0	0	0	1	0	1	0,90309
derita	0	0	0	0	0	0	0	1	1	1	0,90309
kanker	0	0	0	0	0	0	0	1	1	1	0,90309
sayang	0	0	0	0	0	0	0	1	0	1	0,90309
sembuh	0	0	0	0	0	0	0	1	0	1	0,90309
rahasia	0	0	0	1,47712	0	0	0	0	0	1	0,90309
markas	0	0	0	1	0	0	0	0	0	1	0,90309
hancur	0	0	0	1	0	0	0	0	0	1	0,90309
jahat	0	0	0	1	0	0	0	0	0	1	0,90309
pergi	0	0	0	1	0	0	1	0	0	2	0,60206
amerika	0	0	0	1	0	0	0	0	0	1	0,90309
gabung	0	0	0	1	0	0	0	0	0	1	0,90309
rekan	0	0	0	1,30103	0	0	0	0	0	1	0,90309

4.3.2.2 Menghitung TF-IDF Weighting

Proses perhitungan TF-IDF Weighting dilakukan dengan mengkalikan nilai TF dan IDF, proses perhitungan TF-IDF *weighting* dapat menggunakan persamaan (2.3). Berikut contoh untuk melakukan perhitungan TF-IDF pada *term* kisah pada d1.

$$w_{t,d} = w_{tf_{t,d}} * idf_t = 1 * 0{,}20412 = 0{,}20412$$

Pada tabel 4.14 menunjukan hasil perhitungan TF-IDF Weighting.

Tabel 4.14 Hasil TF-IDF *Weighting*

Wtd								
d1	d2	d3	d4	d5	d6	d7	d8	Uji
0,2041	0	0,20412	0	0,2656	0,2041	0	0,2041	0,2041
0,6021	0	0	0	0	0	0	0,7833	0
0,9031	0	0	0	0	0	0	0	0
0,9031	0	0	0	0	0	0	0	0
0,5542	0,426	0	0	0	0	0	0,5542	0
1,1749	0	0	0	0	0	0	0	0
0,6021	0	0	0	0	0	0	0,6021	0
0,9031	0	0	0	0	0	0	0	0
0,9031	0	0	0	0	0	0	0	0
0	0,9031	0	0	0	0	0	0	0
0	0,9031	0	0	0	0	0	0	0
0	0,9031	0	0	0	0	0	0	0
0	0,9031	0	0	0	0	0	0	0
0	0,9031	0	0	0	0	0	0	0
0	0,6021	0,60206	0	0	0	0	0	0
0,5542	0,426	0	0	0	0	0	0,5542	0
0	0	0,60206	0,60206	0	0	0	0	0
0	0	0,90309	0	0	0	0	0	0
0	0	0,90309	0	0	0	0	0	0
0	0	0,60206	0,7833	0	0	0	0	0
0	0	0,90309	0	0	0	0	0	0
0	0	0,90309	0	0	0	0	0	0
0	0	0,90309	0	0	0	0	0	0
0	0	0,90309	0	0	0	0	0	0
0	0	0,90309	0	0	0	0	0	0
0	0	0	0	0,9031	0	0	0	0
0	0	0	0	0,9031	0	0	0	0
0	0	0	0	1,1749	0	0	0	0
0	0	0	0,7833	0,6021	0	0	0	0
0	0	0	0	1,1749	0	0	0	0

0	0	0	0	0,9031	0	0	0	0
0	0	0	0	0,9031	0	0	0	0
0	0	0	0	0,9031	0	0	0	0
0	0	0	0	0,9031	0	0	0	0
0	0	0	0	0	0,7833	0,7833	0	0,7833
0	0	0	0	0	0,6021	0,6021	0	0,6021
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	1,334	0	0	0
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	0,9031	0	0	0
0	0	0	0	0	0	0,6021	0,6021	0
0	0	0	0	0	0	0,9031	0	0
0	0	0	0	0	0	1,8062	0	0
0	0	0	0	0	0	0,9031	0	0
0	0	0	0	0	0	0,9031	0	0
0	0	0	0	0	0	0,9031	0	0
0	0	0	0	0	0	0,9031	0	0
0	0	0	0	0	0	0,9031	0	0
0	0	0	0	0	0	0,9031	0	0,9031
0	0	0	0	0	0	0,7833	0,6021	0
0	0	0	0	0	0	0	0,9031	0
0	0	0	0	0	0	0	1,8062	0
0	0	0	0	0	0	0	1,8062	0
0	0	0	0	0	0	0	0,9031	0
0	0	0	0	0	0	0	0,9031	0
0	0	0	0	0	0	0	0,9031	0
0	0	0	0	0	0	0	0,9031	0
0	0	0	0	0	0	0	1,8062	0
0	0	0	0	0	0	0	0,9031	0
0	0	0	0	0	0	0	0,9031	0
0	0	0	0	0	0	0	0,9031	0,9031
0	0	0	0	0	0	0	0,9031	0,9031
0	0	0	0	0	0	0	0,9031	0
0	0	0	0	0	0	0	0,9031	0
0	0	0	1,33397	0	0	0	0	0
0	0	0	0,90309	0	0	0	0	0
0	0	0	0,90309	0	0	0	0	0

0	0	0	0,90309	0	0	0	0	0
0	0	0	0,60206	0	0	0,6021	0	0
0	0	0	0,90309	0	0	0	0	0
0	0	0	0,90309	0	0	0	0	0
0	0	0	1,17495	0	0	0	0	0

4.3.2.3 Normalisasi

Proses normalisasi dapat dilakukan dengan menggunakan persamaan (2.4). Berikut contoh untuk melakukan perhitungan normalisasi pada *term* kisah pada d1.

$$w_{t,d} = \frac{w_{t,d}}{\sqrt{\sum_{t=1}^{n} {w_{t,d}}^2}} = \frac{0{,}2041}{\sqrt{6{,}023675429}} = \frac{0{,}2041}{2{,}454317711} = 0{,}0832$$

Pada Tabel 4.15 ditunjukan hasil dari perhitungan normalisasi TF-IDF *weighting.*

Tabel 4.15 Hasil Normalisasi TF-IDF *Weighting*

Normalisasi wtd								
d1	d2	d3	d4	d5	d6	d7	d8	Uji
0,0832	0	0,0738	0	0,0934	0,06411	0	0,0444	0,10967
0,2453	0	0	0	0	0	0	0,1705	0
0,368	0	0	0	0	0	0	0	0
0,368	0	0	0	0	0	0	0	0
0,2258	0,19436	0	0	0	0	0	0,1206	0
0,4787	0	0	0	0	0	0	0	0
0,2453	0	0	0	0	0	0	0,131	0
0,368	0	0	0	0	0	0	0	0
0,368	0	0	0	0	0	0	0	0
0	0,41206	0	0	0	0	0	0	0
0	0,41206	0	0	0	0	0	0	0
0	0,41206	0	0	0	0	0	0	0
0	0,41206	0	0	0	0	0	0	0
0	0,41206	0	0	0	0	0	0	0
0	0,27471	0,2176	0	0	0	0	0	0
0,2258	0,19436	0	0	0	0	0	0,1206	0
0	0	0,2176	0,1986	0	0	0	0	0
0	0	0,3264	0	0	0	0	0	0
0	0	0,3264	0	0	0	0	0	0
0	0	0,2176	0,2584	0	0	0	0	0
0	0	0,3264	0	0	0	0	0	0
0	0	0,3264	0	0	0	0	0	0

0	0	0,3264	0	0	0	0	0	0
0	0	0,3264	0	0	0	0	0	0
0	0	0,3264	0	0	0	0	0	0
0	0	0,3264	0	0	0	0	0	0
0	0	0	0	0,3176	0	0	0	0
0	0	0	0	0,3176	0	0	0	0
0	0	0	0	0,4132	0	0	0	0
0	0	0	0,2584	0,2117	0	0	0	0
0	0	0	0	0,4132	0	0	0	0
0	0	0	0	0,3176	0	0	0	0
0	0	0	0	0,3176	0	0	0	0
0	0	0	0	0,3176	0	0	0	0
0	0	0	0	0,3176	0	0	0	0
0	0	0	0	0	0,24602	0,2332	0	0,42084
0	0	0	0	0	0,18909	0,1792	0	0,32346
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0,41897	0	0	0
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0,28364	0	0	0
0	0	0	0	0	0	0,1792	0,131	0
0	0	0	0	0	0	0,2688	0	0
0	0	0	0	0	0	0,5376	0	0
0	0	0	0	0	0	0,2688	0	0
0	0	0	0	0	0	0,2688	0	0
0	0	0	0	0	0	0,2688	0	0
0	0	0	0	0	0	0,2688	0	0
0	0	0	0	0	0	0,2688	0	0
0	0	0	0	0	0	0,2688	0	0,48519
0	0	0	0	0	0	0,2332	0,131	0
0	0	0	0	0	0	0	0,1965	0
0	0	0	0	0	0	0	0,3931	0
0	0	0	0	0	0	0	0,3931	0
0	0	0	0	0	0	0	0,1965	0
0	0	0	0	0	0	0	0,1965	0
0	0	0	0	0	0	0	0,1965	0
0	0	0	0	0	0	0	0,1965	0
0	0	0	0	0	0	0	0,3931	0

0	0	0	0	0	0	0	0,1965	0
0	0	0	0	0	0	0	0,1965	0
0	0	0	0	0	0	0	0,1965	0,48519
0	0	0	0	0	0	0	0,1965	0,48519
0	0	0	0	0	0	0	0,1965	0
0	0	0	0	0	0	0	0,1965	0
0	0	0	0,44	0	0	0	0	0
0	0	0	0,2979	0	0	0	0	0
0	0	0	0,2979	0	0	0	0	0
0	0	0	0,2979	0	0	0	0	0
0	0	0	0,1986	0	0	0,1792	0	0
0	0	0	0,2979	0	0	0	0	0
0	0	0	0,2979	0	0	0	0	0
0	0	0	0,3876	0	0	0	0	0

4.3.3 Cosine Similarity

Perhitungan *cosine similiraty* dihitung dengan mengkalikan data latih dengan data uji yang ingin diketahui kategorinya. Adapun persamaan yang digunakan untuk menghitung *cosine similarity* dapat dilihat pada persamaan (2.5). Berikut contoh perhitungan *cosine similarity term* kisah pada d1.

$$Cosine\ (d_i, q_i) = 0{,}08316771 * 0{,}10966569 = \ 0{,}009121$$

Pada Tabel 4.16 ditunjukan hasil perhitungan nilai *cosine similarity*.

Tabel 4.16 Hasil *Cosine Similarity*

CosSim							
(d1 * uji)	**(d2 * uji)**	**(d3 * uji)**	**(d4 * uji)**	**(d5 * uji)**	**(d6 * uji)**	**(d7 * uji)**	**(d8 * uji)**
0,009121	0	0,00809	0	0,01024	0,00703	0	0,004871
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0,10353	0,09812	0
0	0	0	0	0	0,06117	0,05797	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0,13043	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0,095357
	0	0	0	0	0	0	0	0,095357
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
Jumlah	0,009121	0	0,00809	0	0,01024	0,17173	0,28652	0,195585

Setelah didapatkan hasil *cosine similarity* maka dilakukan pengurutan tingkat kemiripan dari yang terbesar hingga yang terkecil. Tabel 4.17 menunjukan urutan tingkat kemiripan terhadap data uji.

Tabel 4.17 Urutan Kemiripan Data Uji

Dokumen	Nilai	Kategori
7	0,28652	Keluarga
8	0,195585	Keluarga
6	0,17173	Thriller
5	0,01024	Horor
1	0,009121	Romantis
3	0,00809	Aksi
2	0	Romantis
4	0	Aksi

4.3.4 Klasifikasi dengan Improved K-NN

Proses pengklasifikasian dimulai dengan menentukan nilai n (nilai *k* baru) yang dapat menggunakan persamaan (2.6). Untuk nilai *k* awal ditetapkan dengan nilai 4 untuk tiap kategori. Tabel 4.18 menunjukan banyak data latih.

Tabel 4.18 Banyak Data Latih

Data latih					
Romantis	**Aksi**	**Horor**	**Thriler**	**Keluarga**	**Total**
2	2	1	1	2	8

Setelah menghitung nilai *n*, maka didapatkan nilai *k* baru. Berikut contoh melakukan perhitungan *n* pada kategori romantis.

$$n = \left[\frac{4 * 2}{2}\right] = 4$$

Tabel 4.19 menunjukan hasil dari nilai *n* (nilai *k* baru).

Tabel 4.19 Nilai *n*

	N				
K-Values	**Romantis**	**Aksi**	**Horor**	**Thriler**	**Keluarga**
4	4	4	2	2	4

Setelah mendapatkan nilai n, maka selanjutnya dilakukan perhitungan probabilitas dokumen uji pada tiap kategori dengan menggunakan persamaan 2.7. Berikut contoh menghitung kemungkinan pada kategori romantis.

$$p(x, c_{Romantis}) = \frac{((0{,}29124 * 0) + (0{,}19559 * 0) + (0{,}17173 * 0) + (0{,}01025 * 0))}{(0{,}29124 + 0{,}19559 + 0{,}17173 * 0 + 0{,}01025 * 0)}$$

Dari hasi perhitungan dengan menggunakan persamaan (2.7) maka didapatkan hasil seperti yang ditunjukan oleh Tabel 4.20

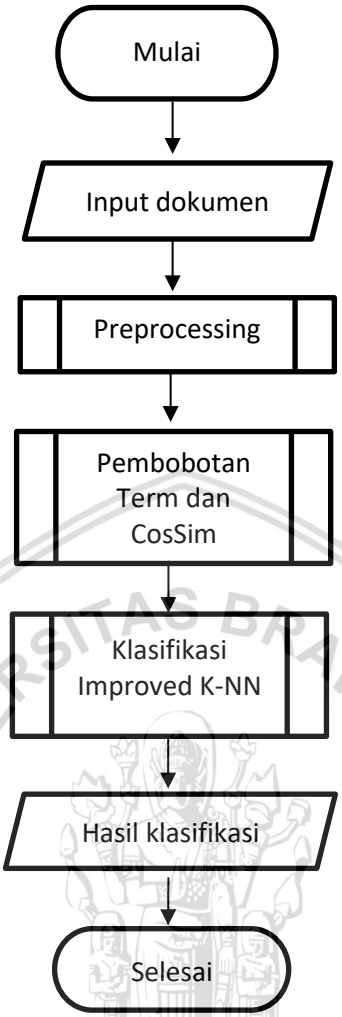
Tabel 4.20 Hasil Klasifikasi

Romantis	**Aksi**	**Horor**	***Thriller***	**Keluarga**
0	0	0	0	0,725982

Hasil perhitungan probabilitas menunjukan bahwa nilai tertinggi ditunjukan oleh kategori keluarga sebesar 0,725982.

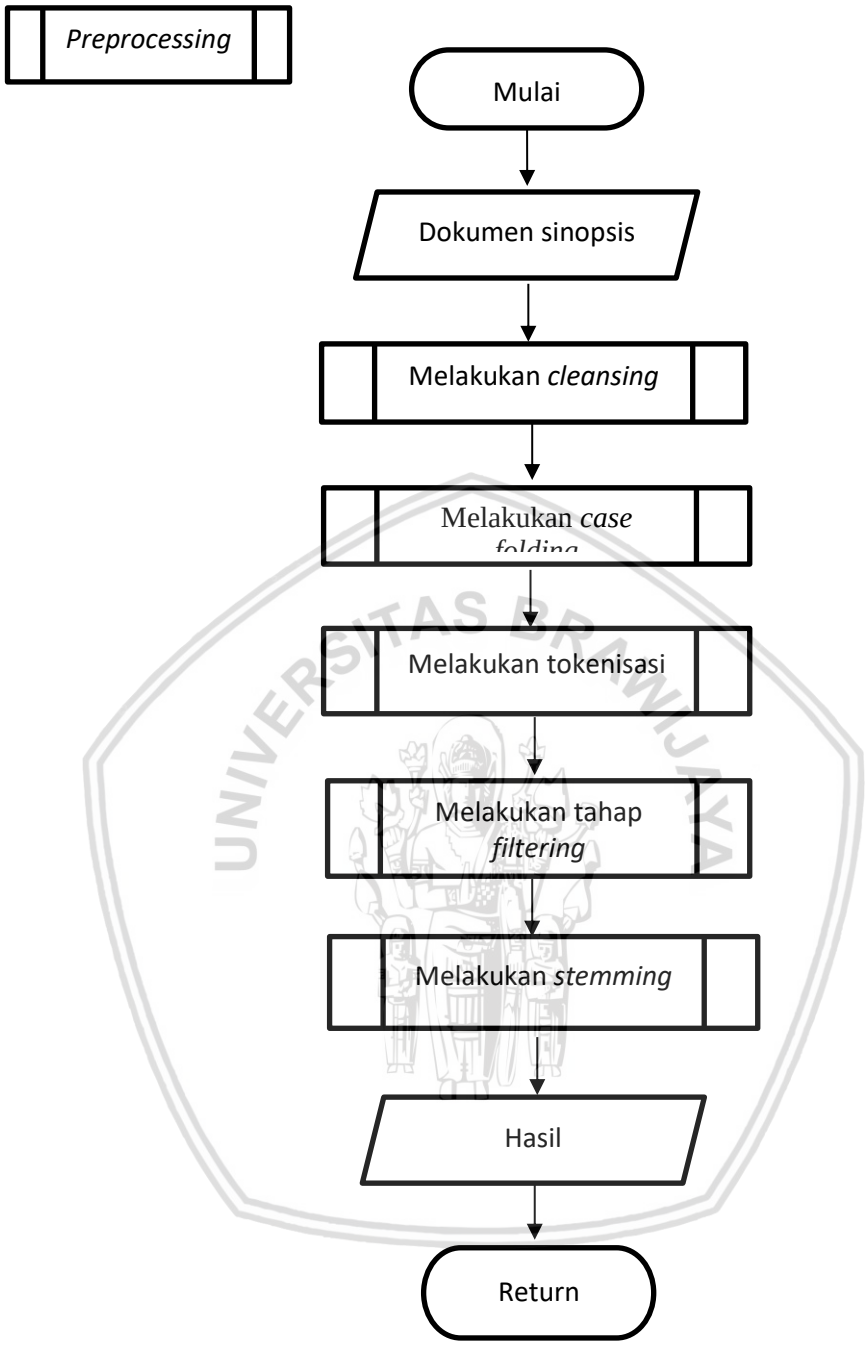
4.4 Diagram Alir Sistem

Pada Gambar 4.1 dijelaskan gambaran umum tentang sistem yang akan diberikan *input* dokumen berupa sinopsis film yang kemudian akan diproses. Dimulai dengan melakukan *preprocessing*, pembobotan *term* hingga normalisasi, perhitungan *cosine similarity*, dan melakukan klasifikasi pada hasil perhitungan *cosine similarity* dengan menggunakan metode Improved K-NN. Dimana *output* yang akan dihasilkan ialah berupa kategori romantis, aksi, horor, *thriller*, atau keluarga sesuai dengan hasil perhitungan dan pengklasifikasian data uji.



Gambar 4.1 Diagram Alir Sistem

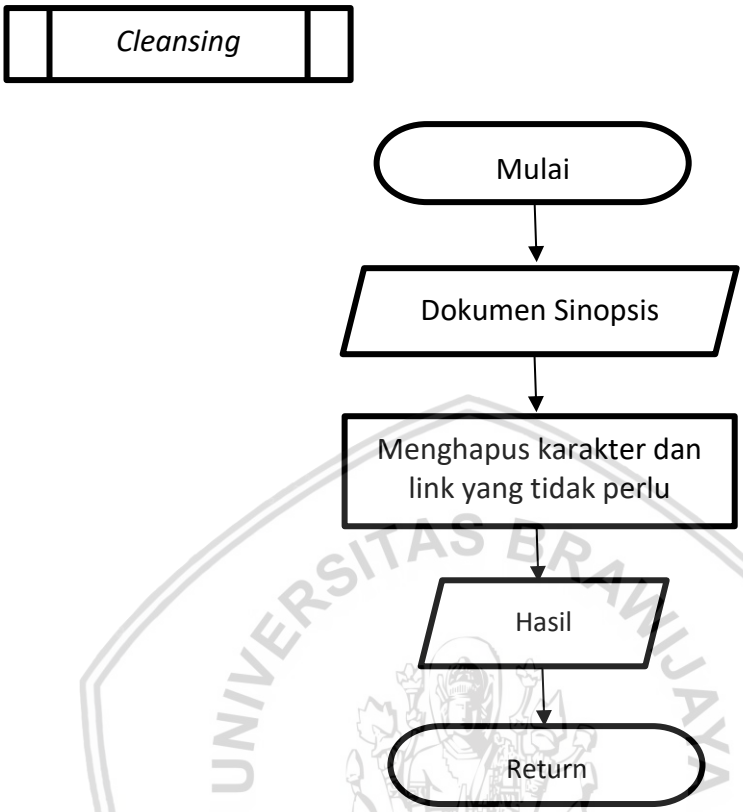
Pada Gambar 4.1 menunjukan diagram alir dari sistem, proses dimulai dengan memasukan dokumen uji berupa sinopsis film kemudian dokumen tersebut akan melewati tahapan *preprocessing* dimana tahapan ini dilakukan dengan beberapa proses yakni *cleansing*, *case folding*, tokenisasi, *filtering*, dan *stemming*. Kemudian data uji yang telah di *preprocessing* akan melewati proses pembobotan *term* yang memiliki beberapa tahapan yakni menghitung nilai TF kemudian melakukan pembobotan pada TF, lalu menghitung nilai TF-IDF setelah itu melakukan normalisasi pada hasil pembobotan. Kemudian hasil pembobotan *term* akan dihitung untuk mendapatkan nilai *cosine similarity.* Hasil dari *cosine similarity* akan diurutkan secara menurun kemudian dilakukan proses pengklasifikasian dari hasil *cosine similarity* tersebut dengan menggunakan metode Improved K-NN. Setelah dilakukan klasifikasi maka akan didapatkan hasil berupa kategori yang sesuai untuk data yang diujikan.



Gambar 4.2 Diagram Alir *Preprocessing*

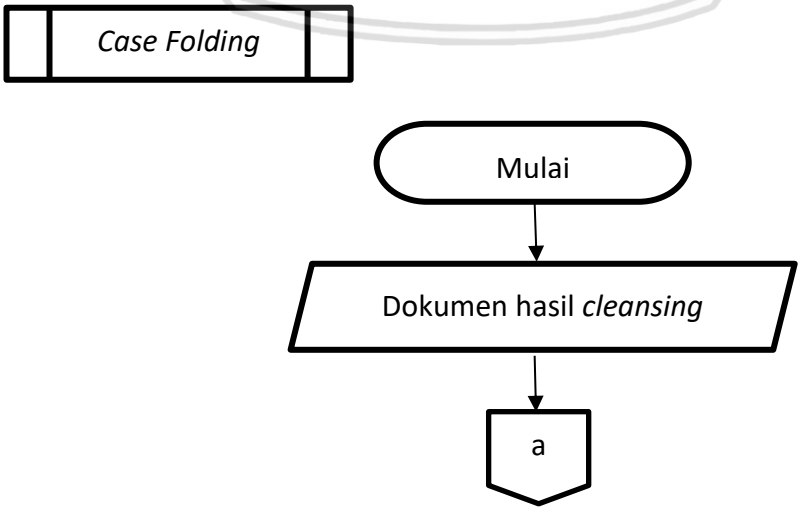
Pada Gambar 4.2 ditunjukan diagram alir dari proses *preprocessing*, mulai dari melakukan proses *cleansing* dimana membersihkan *text* pada dokumen dari karakter yang tidak perlu serta link-link, lalu melakukan tahap *case folding* yakni mengubah semua huruf menjadi huruf kecil, kemudian melakukan tokenisasi untuk menghapus semua memisah kata menjadi token-token, selanjutnya proses *filtering* dimana dilakukan penghapusan kata yang *stopword*, terakhir melakukan tahap *stemming* dimana melakukan perubahan pada kata sehingga menjadi kata dasar. Hasil dari tahapan-tahapan yang ada pada *preprocessing* ialah berupa

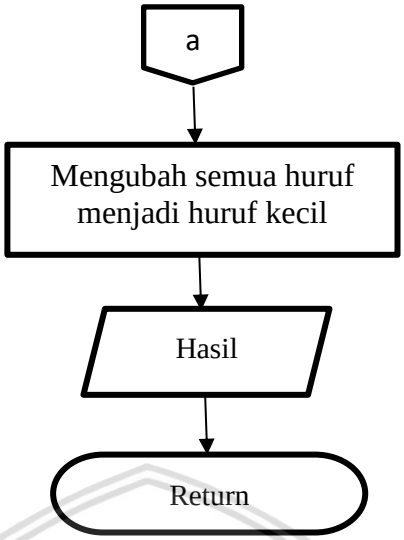
daftar kata (*term*) yang kemudian akan diproses untuk melakukan pembobotan *term* dan klasifikasi dengan menggunakan metode Improved K-NN.



Gambar 4.3 Diagram Alir *Cleansing*

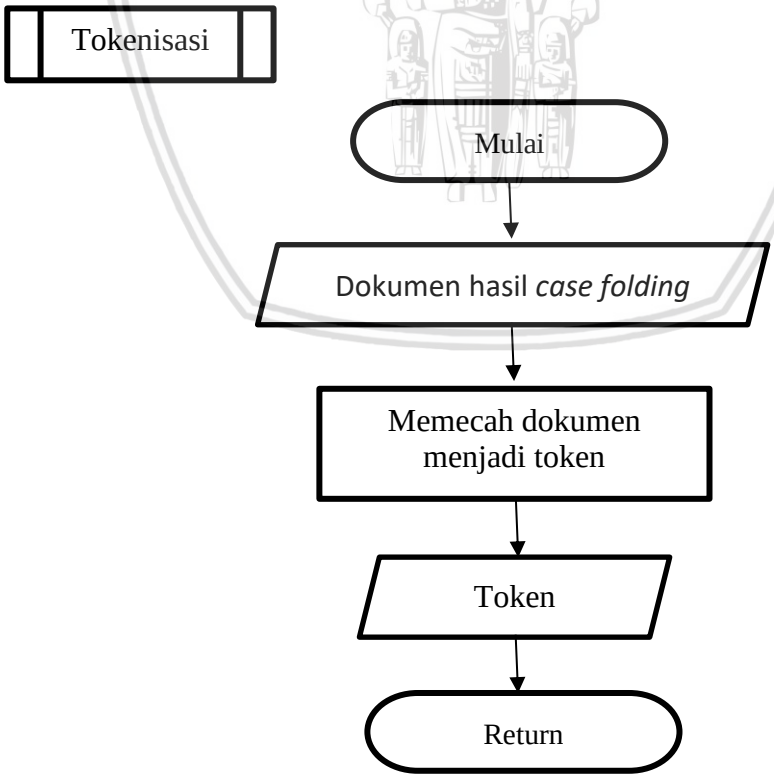
Pada Gambar 4.3 ditunjukkan diagram alir dari proses *cleansing*, proses dimulai dari memasukan dokumen berupa sinopsis kemudian dilakukan proses penghapusan karakter-karakter yang tidak perlu serta link-link, selanjutnya akan didapatkan hasil berupa dokumen yang telah dibersihkan dari karakter-karakter yang tidak perlu dan link-link.





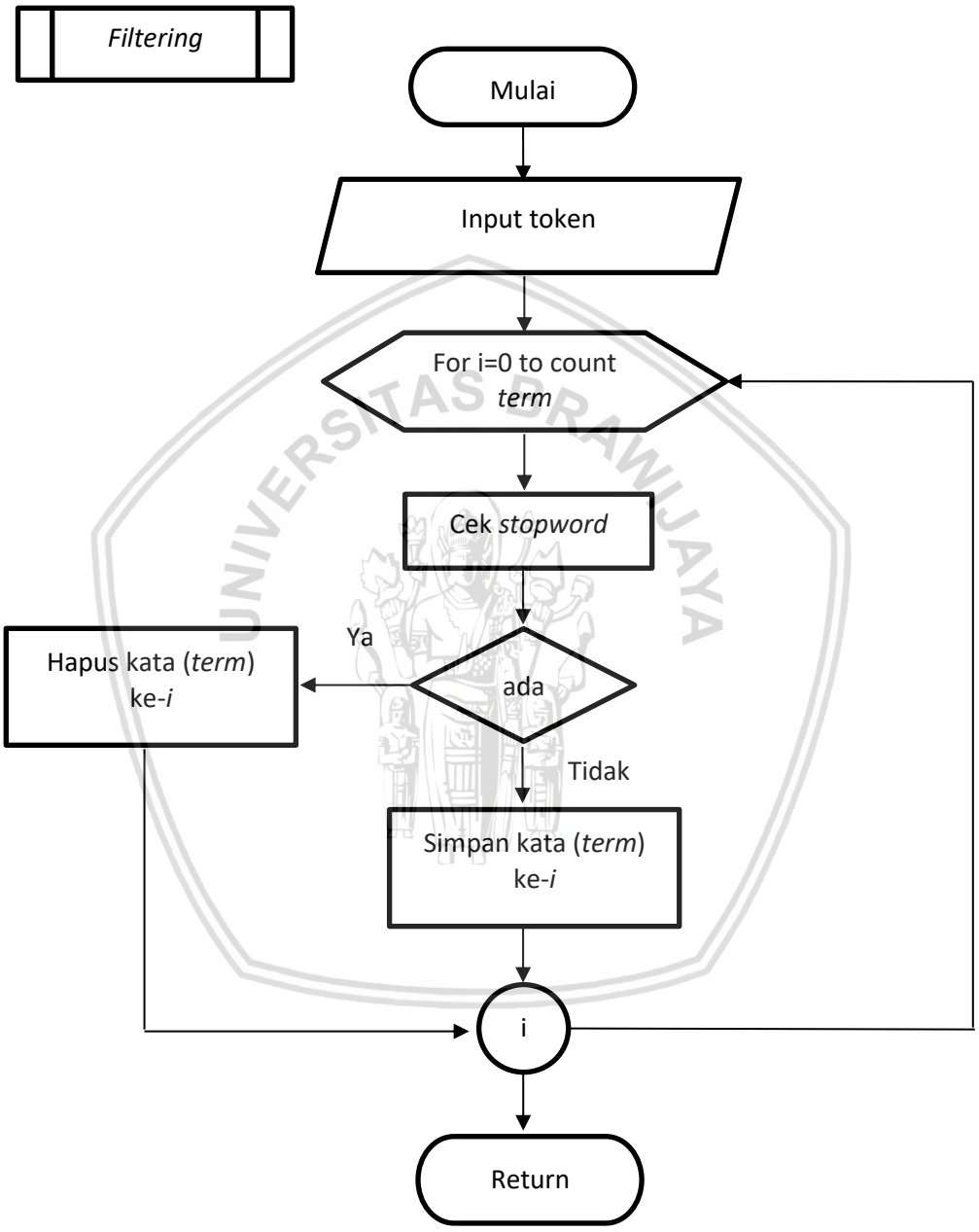
Gambar 4.4 Diagram Alir *Case Folding*

Pada Gambar 4.4 ditunjukkan diagram alir dari proses *case folding*, proses dimulai dari memasukan dokumen yang telah melewati tahap *cleansing* kemudian dilakukan proses mengubah seluruh huruf pada dokumen menjadi huruf kecil, selanjutnya akan didapatkan hasil berupa dokumen yang seluruh isinya berupa huruf kecil.



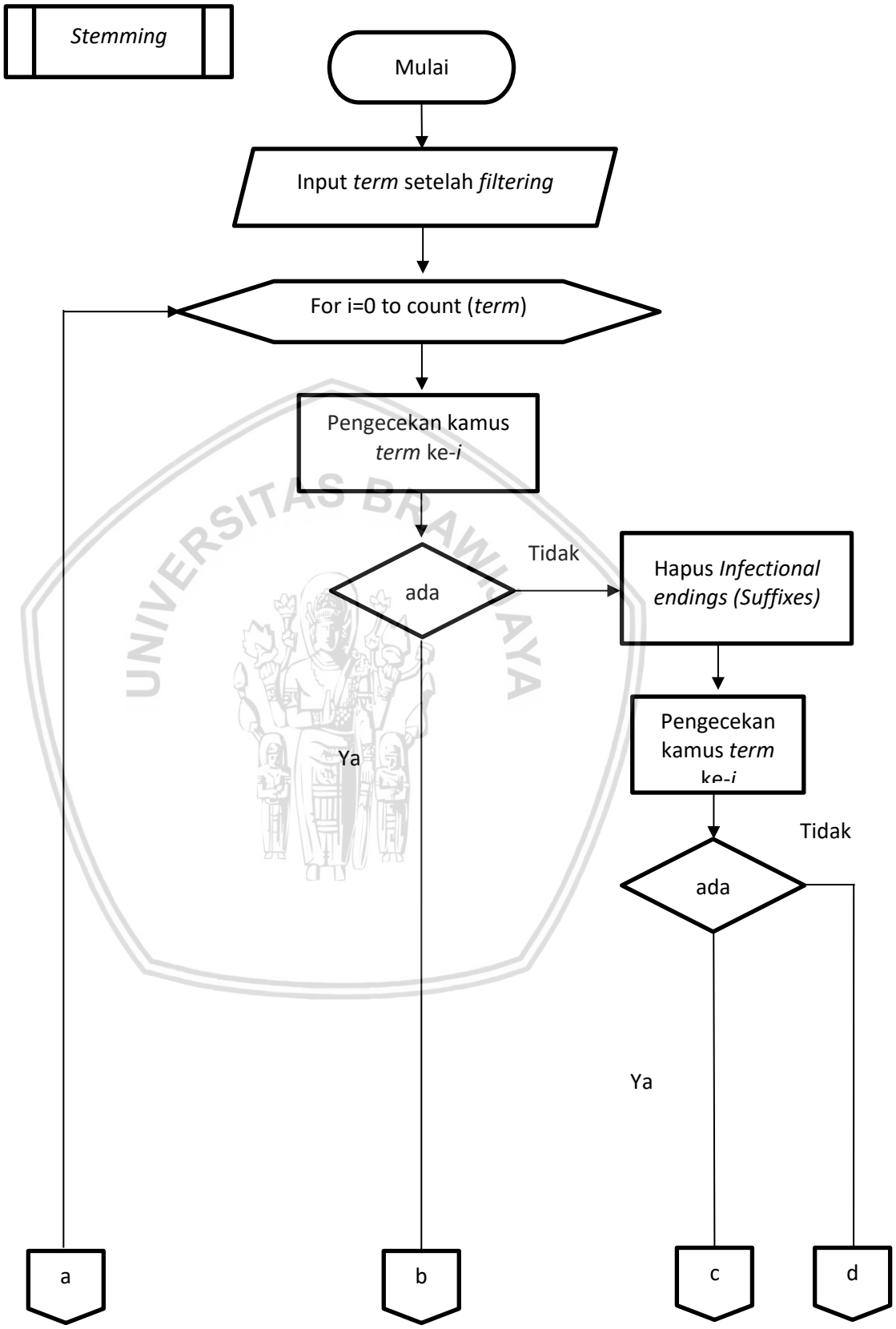
Gambar 4.5 Diagram Alir Tokenisasi

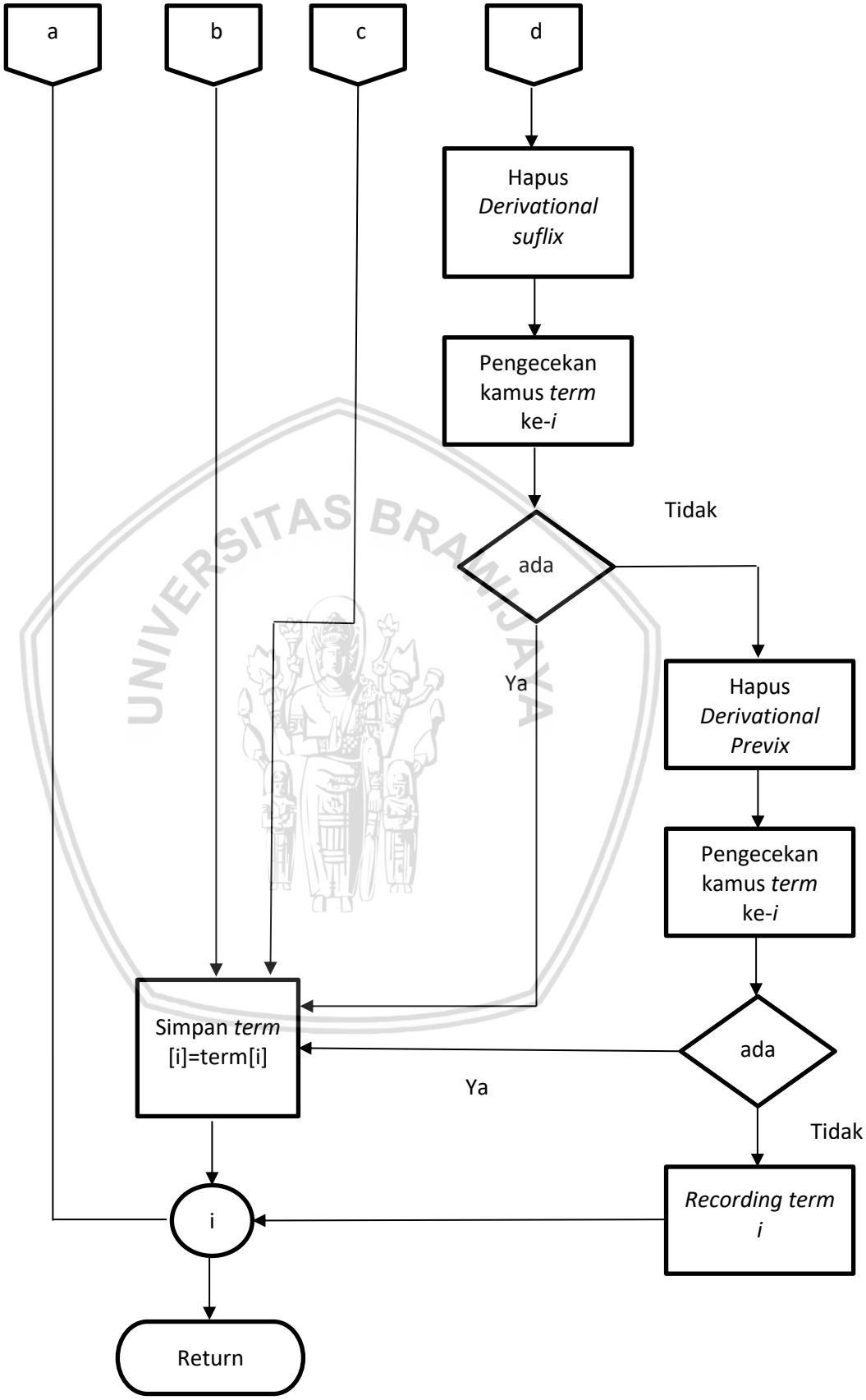
Pada Gambar 4.5 ditunjukkan diagram alir dari proses tokenisasi, proses dimulai dari memasukan dokumen yang telah melewati tahap *case folding* kemudian dilakukan proses untuk memisahkan string pada dokumen atau memecah dokumen menjadi token, selanjutnya akan didapatkan hasil berupa dokumen yang telah menjadi token.



Gambar 4.6 Diagram Alir *Filtering*

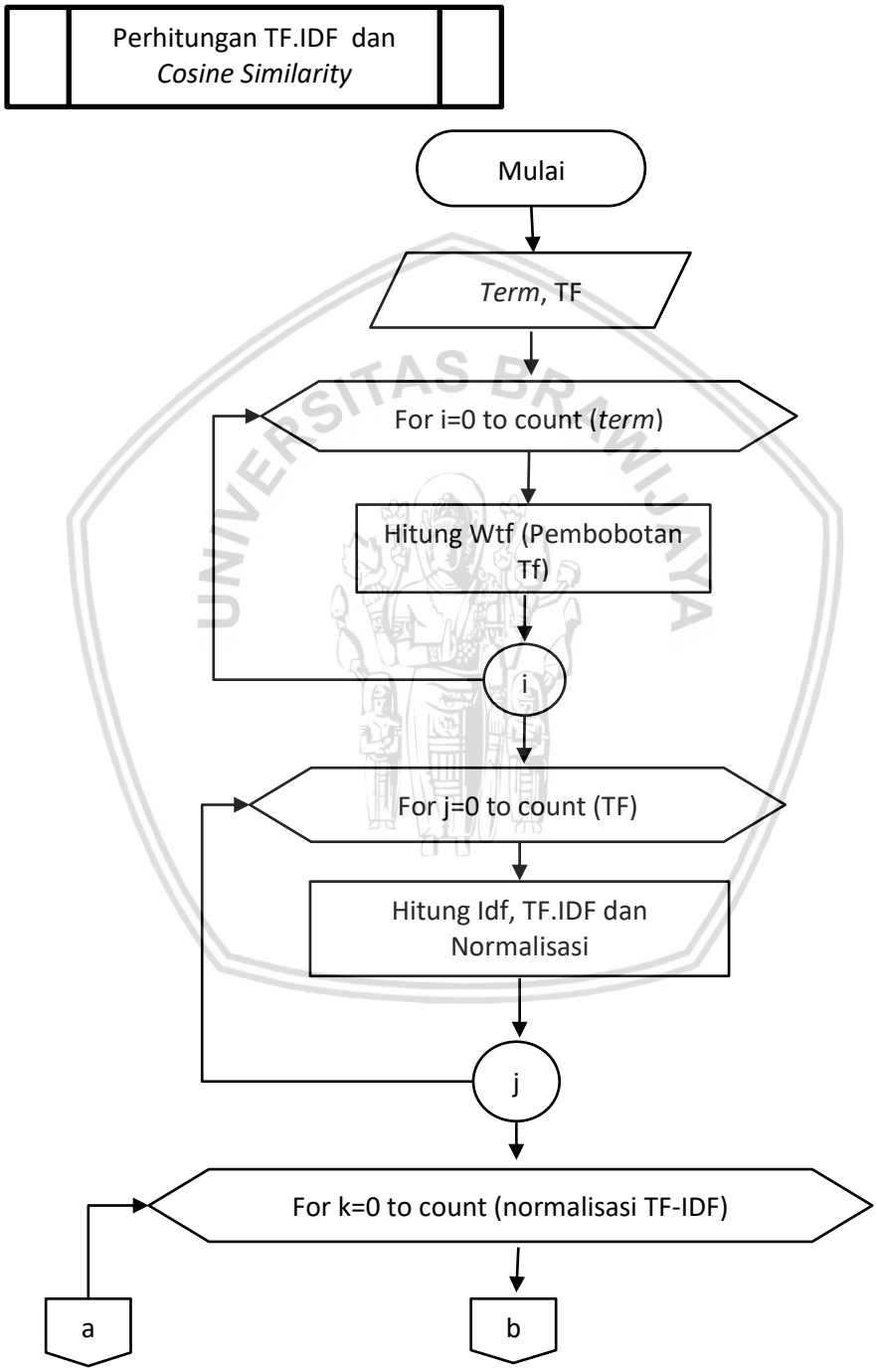
Pada Gambar 4.6 ditunjukkan diagram alir dari proses *filtering*, proses dimulai dari memasukan kata kemudian akan dicocokan dengan daftar kata dalam *stopword*. Apabila kata terdapat dalam *stopword* maka kata akan dihapus, namun sebaliknya jika kata tidak terdaftar maka kata akan disimpan pada basis data dan dilanjutkan keproses selanjutnya.

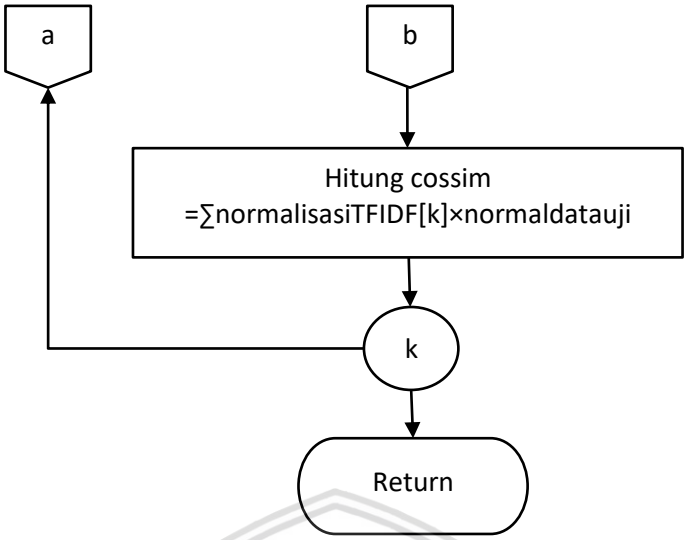

Stemming
Mulai
Input *term* setelah *filtering*
For i=0 to count (*term*)
Pengecekan kamus *term* ke-*i*
ada
Tidak
Hapus *Infectional endings (Suffixes)*
Ya
Pengecekan kamus *term* ke-*i*
Tidak
ada
Ya
a
b
c
d



Gambar 4.7 Diagram Alir *Stemming*

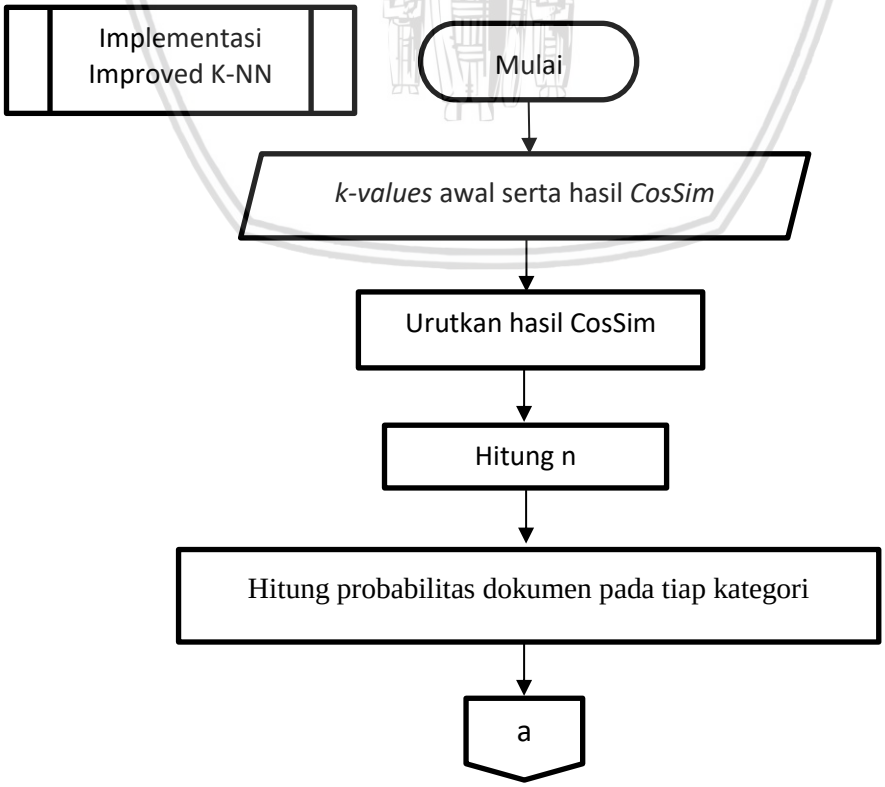
Pada Gambar 4.7 ditunjukan diagram alir dari proses *stemming*, dimana algoritma yang digunakan ialah algoritma *stemming* Nazief dan Andriani. Proses dimulai dari memasukan kata kemudian menghilangkan *inflectional endings*, *derivational suffix dan prefix,* serta melakukan pengecekan pada daftar kata dasar yang akan digunakan.

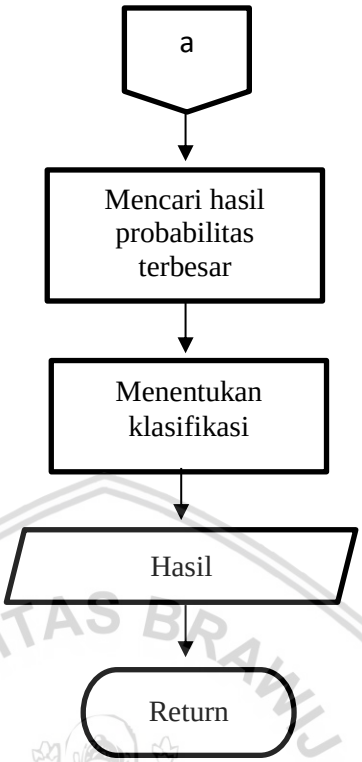




Gambar 4.8 Diagram Alir TF-IDF dan *Cosine Similarity*

Gambar 4.8 ditunjukan diagram alir proses pembobotan *term* (kata), dimulai dengan memasukan seluruh *term* yang telah melalui *preprocessing* lalu dilakukan pembobotan *term*, yaitu dengan menghitung nilai Wtf dari hasil nilai TF. Kemudian melakukan peritungan nilai TF-IDF, nilai IDF didapatkan dengan melakukan perkalian pada dokumen latih dan hasil WTF. Lalu hasil dari perhitungan TF-IDF dinormalisasi. Setelah didapatkan hasil dari normalisasi kemudian dilanjutkan ke proses cosine similarity pada data uji terhadap data latih untuk mengetahui hasil kemiripan antara data uji dan data latih.





Gambar 4.9 Diagram Alir Klasifikasi Improved K-NN

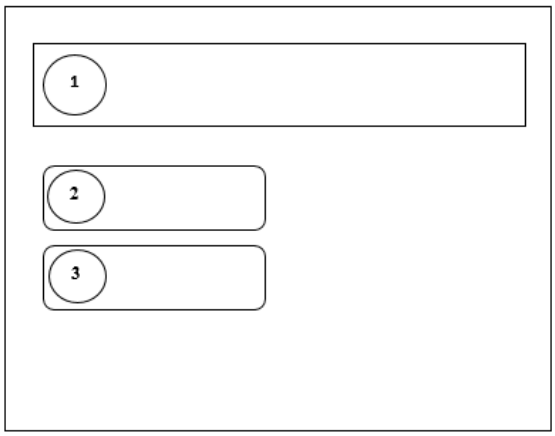
Gambar 4.9 menunjukan diagram alir dari proses klasifikasi menggunakan metode Improved K-NN. Proses klasifikasi dimulai dengan melakukan input nilai *k* awal serta hasil dari perhitungan cosine similarity, setelah itu mengurutkan hasil dari cosine similarity yang telah didapatkan sebelumnya. Selanjutnya melakukan perhitungan nilai n dilakukan pada tiap kategori yang ada sesuai dengan nilai *k* awal. Lalu menghitung probabilitas dan melakukan pencarian probabilitas tertinggi, nilai probabilitas tertinggi merupakan hasil kategori untuk dokumen yang diujikan.

4.5 Perancangan Antarmuka (*User Interface*)

Perancancang antarmuka (UI) dibutuhkan sebagai penghubung antara pengguna dengan sistem. Adapun perancangan antarmuka yang akan dibangun dalam sistem ialah sebagai berikut :

4.5.1 Halaman Awal

Halaman awal ialah halaman utama yang menampilkan judul dari sistem, pada halaman awal ini terdapat dua buah tombol. Pada tombol pertama digunakan untuk melakukan proses pengujian pada data uji dengan menggunakan data latih dan pada tombol kedua digunakan untuk pengguna dan berfunsi pula untuk mengalihkan pengguna kehalaman selanjutnya yakni halaman pengguna. Tampilan halaman awal dapat dilihat pada Gambar 4.10.



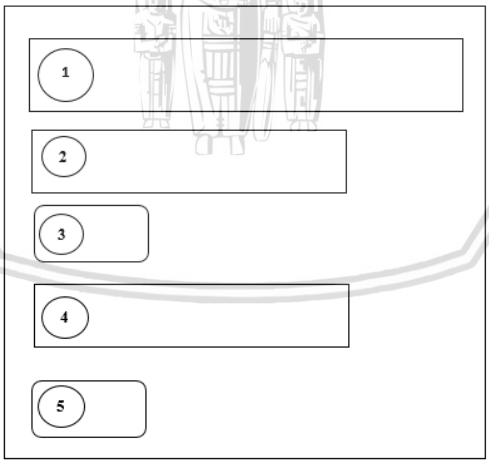
Gambar 4.10 Halaman Awal

Keterangan :

1. Judul sistem
2. Tombol melakukan klasifikasi atau pengujian
3. Tombol untuk menuju ke halaman pengguna

4.5.2 Antarmuka Pengujian

Pada antarmuka pengujian berfungsi untuk memasukan data uji berupa sinopsis film, sehingga data uji dapat diproses dan diklasifikasikan sesuai kategorinya.



Gambar 4.11 Halaman Pengujian

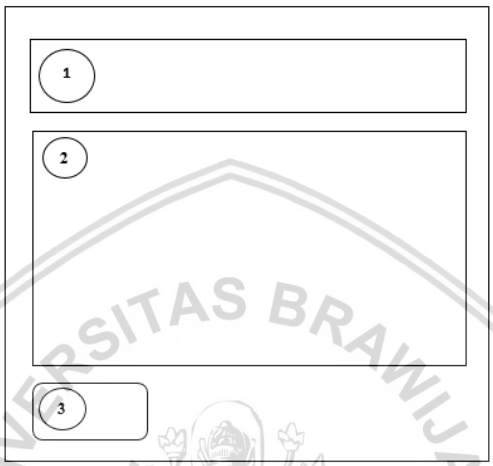
Keterangan :

1. Judul sistem
2. Kolom untuk mengisi data uji dan untuk memilih kategori awal
3. Tombol untuk *input*
4. Kolom berisi keterangan kategori

5. Tombol untuk kembali ke halaman awal

4.5.3 Halaman Hasil Pengujian

Pada halaman hasil pengujian akan menampilkan hasil dari pengujian pada data uji yang telah dilakukan sebelumnya. Sebelum memunculkan hasil kalsifikasi, terlebih dahulu data uji yang dimasukan melewati beberapa proses yakni *preprocessing*, pembobotan, hingga mendapatkan hasil kategori yang sesuai dengan data uji.



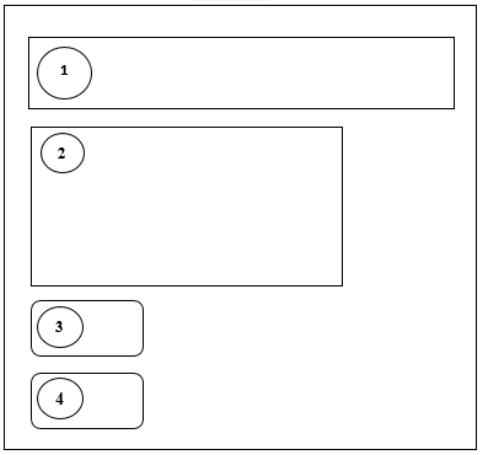
Gambar 4.12 Halaman Hasil Pengujian

Keterangan :

1. Judul sistem
2. Kolom berisi hasil pengujian
3. Tombol kembali

4.5.4 Halaman Pengguna

Pada halaman ini akan menampilkan antar muka untuk pengguna memasukan data berupa sinopsis film untuk dilakukan klasifikasi.



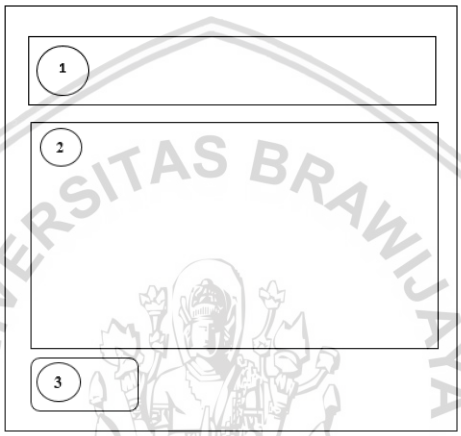
Gambar 4.13 Halaman Pengguna

Keterangan :

1. Judul Sistem
2. Kolom untuk memasukan sinopsis film
3. Tombol untuk *input* data
4. Tombol kembali kehalaman utama

4.5.5 Halaman Hasil Klasifikasi

Pada halaman ini akan menampilkan hasil dari pengujian pada data yang di masukan oleh pengguna dan akan menampilkan hasil dari proses *preprocessing* dan kategori untuk data yang diujikan.



Gambar 4.14 Halaman Hasil Klasifikasi

Keterangan :

1. Judul Sistem
2. Kolom berisi hasil pengujian
3. Tombol kembali kehalaman utama

4.6 Perancangan *Database*

Perancangan *database* atau basis data dibangun untuk melakukan penyimpanan data, hasil dari proses *preprocessing*, perhitungan, hingga hasil klasifikasi.

4.6.1 Tabel Data Latih

Tabel data latih merupakan tabel untuk melakukan penyimpanan informasi yang berkaitan dengan data latih yang digunakan. Tabel 4.21 menunjukan struktur yang dimiliki oleh tabel data latih.

Tabel 4.21 Struktur Tabel Data Latih

No	Nama Field	Type	Size
1	Id	Int	10

2	Kata	Varchar	255
3	d0	Int	255
4	d1	Int	255
5	d2	Int	255
6	d3	Int	255
7	d4	Int	255
8	d5	Int	255
9	d6	Int	255
10	d7	Int	255

4.6.2 Tabel Normalisasi

Tabel normalisasi merupakan tabel untuk melakukan penyimpanan informasi yang berkaitan dengan normalisasi data. Tabel 4.22 menunjukan struktur yang dimiliki oleh tabel normalisasi.

Tabel 4.22 Struktur Tabel Normalisasi

No	Nama Field	Type	Size
1	Id	Int	10
2	Kata	Varchar	255
3	d0	Float	255
4	d1	Float	
5	d2	Float	
6	d3	Float	
7	d4	Float	
8	d5	Float	
9	d6	Float	
10	d7	Float	

4.7 Perancangan Pengujian dan Analisis

Perancangan pada pengujian dilakukan untuk mengetahui adanya kesalahan atau tidak ketika melakukan implementasi dengan menggunakan metode Improved K-NN. Tabel *confusion matrix* dapat mempermudah proses untuk melakukan perhitungan pada *precision*, *recall*, dan *f1-measure* serta akurasi. Dengan begitu akan diketahui faktor-faktor yang memberikan pengaruh ketepatan pada hasil klasifikasi dengan metode Improved K-NN. Pengujian akan dilakukan dengan beberapa skenario, dimana jumlah dari data latih tidak lah

sama atau berbeda-beda. Tabel 4.23 menunjukan perancangan dari tabel skenario.

Tabel 4.23 Perancangan Tabel Skenario

Skenario	Data Latih						Data Uji					
	K1	K2	K3	K4	K5	Jumlah	K1	K2	K3	K4	K5	Jumlah

Keterangan :

- K1 = Romantis
- K2 = Aksi
- K3 = Horor
- K4 = *Thriller*
- K5 = Keluarga

Pengujian pada tiap skenario akan dilakukan menggunakan nilai *k* awal yang memiliki nilai yang berbeda-beda. Pada Tabel 4.24 ditujukan perancangan tabel pengujian.

Tabel 4.24 Perancangan Tabel Pengujian

Nilai *k*	*n*					Pengujian			
	Romantis	Aksi	Horor	*Thriller*	Keluarga	*Precision*	*Recall*	*F-Measure*	Akurasi

4.8 Kesimpulan

Kesimpulan dihasilkan setelah keseluruhan proses dalam pengklasifikasian telah dilakukan. Kesimpulan didapatkan dari hasil analisa pengujian yang telah dilakukan sehingga dapat menghasilkan saran yang diharapkan dapat berguna dalam memperbaiki kekurangan penelitian selanjutnya.

4.9 Spesifikasi Sistem

Untuk membuat sistem yang memiliki fungsi sesuai dengan kebutuhan yang dibutuhkan maka pengimplementasian sistem mengacu pada proses dan hasil

analisis kebutuhan dan perancangan yang telah dibahas pada bab sebelumnya. Spesifikasi sistem dibagi menjadi dua yakni spesifikasi *hardware* (perangkat keras) dan *software* (perangkat lunak).

4.9.1 Spesifikasi Perangkat Keras

Pada pengembangan sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN, memanfaatkan komputer dengan spesifikasi *hardware* (perangkat keras) sebagai berikut :

a. Processor Intel Pentium 987 1.5GHz 2MB
b. Kapasitas Memori (RAM) 2.00 GB

4.9.2 Spesifikasi Perangkat Lunak

Pada pengembangan sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN, memanfaatkan komputer dengan spesifikasi *software* (perangkat lunak) sebagai berikut :

a. OS Windows 10 Profesional 32 bit
b. Bahasa Pemrograman PHP dan HTML version : 5.6.31
c. XAMPP v3.2.2
d. Notepad ++ Text Editor
e. Materialize (Front-end framework)
f. Google Chrome Versi 66.0.3359.181

4.10 Batasan Implementasi

Batasan implementasi ialah batasan-batasan yang dimiliki oleh sistem hal ini mengacu pada seberapa banyak proses yang mampu dilakukan oleh sistem yang mana hal ini sesuai dengan perancangan sistem yang dibangun. Adapun batasan implementasi bermanfaat untuk memberikan penjelasan tentang ruang lingkup pengimplementasian sistem. Beberapa batasan implementasi sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN ialah sebagai berikut :

1. Sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN dirancang serta dijalankan dengan memanfaatkan aplikasi berbasis web.
2. Metode yang di manfaatkan ialah metode Improves K-Nearest Neighbor (K-NN).
3. Dokumen untuk data lati dan data uji diperoleh dari beberapa situs online yakni sinopsisfilem21.com, posfilm.com, filmbioskop.co.id, pusatsinopsis.com, filmbioskop.net, hype.idntimes.com, filmbor.com, sinopsisfilm.co.id, sinopsisdanreviewfilm.blogspot.com, dan industry.co.id.

4. *Output* yang dihasilkan ialah hasil klasifikasi yang berupa kategori film yakni romantis, horor, aksi, keluarga dan *thriller*.

Penentuan klasifikasi didasarkan dari frekuensi kemunculan kata yang terdapat pada dokumen.

4.11 Implementasi

Sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN memiliki beberapa proses yang terdiri dari *preprocessing*, pembobotan *term* (*term weighting*) serta klasifikasi dengan menggunakan metode improve K-NN. Data yang digunakan sebagai masukan untuk melakukan pengujian ialah berupa sinopsis film sebagai dokumen yang kemudian akan di olah sehingga dapat menghasilkan keluaran yang berupa kategori genre film romantis, aksi, horor, *thriller*, dan keluarga terhadapa sinopsis film yang dimasukan.

4.11.1 Preprocessing

Proses *preprocessing* memiliki beberapa tahapan yakni *cleansing, case folding,* tokenisasi, *filtering,* dan, *stemming.*

4.11.1.1 Cleansing

Pada tahap ini dilakukan penghapusan URL serta angka, karakter selain huruf dan tanda baca. Implementasi dari tahap *cleansing* dapat dilihat pada *Source Code* 5.1.

```
1  $hapussimbol ="/[^a-zA-Z]/";
2  $cleansing = preg_replace($hapussimbol, ' ', $data);
```

Source Code* 4.1 Implementasi Tahap *Cleansing

Berikut ini merupakan penjelasan *Source Code* dari Implementasi *cleansing* yang ditunjukkan pada Tabel 4.25.

Tabel 4.25 Penjelasan *Source Code Cleansing*

1	Berisi karakter yang ingin dihilangkan
2	Fungsi `preg_replace`, digunakan untuk mengganti karakter yang ada pada dokumen dengan cara menghapus karakter-karakter tersebut

4.11.1.2 Case Folding

Pada tahap ini dilakukan pengubahan data, yakni mengubah semua huruf pada dokumen menjadi huruf kecil (*lower case*). Implementasi dari tahap *case folding* dapat dilihat pada *Source Code* 4.2.

```
1  $casefolding = strtolower($cleansing);
```

Source Code* 4.2 Implementasi Tahap *Case Folding

Berikut ini merupakan penjelasan *Source Code* dari Implementasi *case folding* yang ditunjukkan pada Tabel 4.26.

Tabel 4.26 Penjelasan *Source Code Case Folding*

1	`strtolower` berfungsi untuk melakukan perubahan pada data yang sudah melewati tahap *cleansing* menjadi huruf kecil

4.11.1.3 Tokenisasi

Pada tahap ini dilakukan pemisahan pada setiap kata yang dipisahkan oleh *whitespace* sehingga menjadi token, hal ini dilakukan pada semua dokumen. Implementasi dari tahap tokenisasi dapat dilihat pada *Source code* 4.3.

```
1  $tokenisasi = explode(' ', $casefolding);
2  $tokenisasi = array_filter($tokenisasi);
3  sort($tokenisasi);
```

***Source Code* 4.3 Implementasi Tahap Tokenisasi**

Berikut ini merupakan penjelasan *Source Code* dari Implementasi tokenisasi yang ditunjukkan pada Tabel 4.27.

Tabel 4.27 Penjelasan *Source Code* Tokenisasi

1	Memisahkan string yang dipecah oleh *whitespace*
2	Menyaring array yang sesuai dengan fungsi `$tokenisasi`

4.11.1.4 Filtering

Pada tahap ini dilakukan penghapusan pada kata yang dirasa kurang penting yang termasuk pada *stoplist*, seperti itu, adalah, di, ke, ini, saya, aku, kamu dan lain sebagainya. Implementasi dari tahap *filtering* dapat dilihat pada *Source Code* 4.4.

```
1  $selectstoplist = "05-stopword_list.txt";
2  $read = fopen($selectstoplist, "r");
3  $readstoplist = fread($read, filesize($selectstoplist));
4  $daftarstoplist = explode("\n", $readstoplist);
5  fclose($read);
6  $daftarstoplist = array_filter($daftarstoplist);
7  $filtering = array_values(array_diff(array_map("trim",
   $tokenisasi), array_map("trim", $daftarstoplist)));
```

Source Code* 4.4 Implementasi Tahap *Filtering

Berikut ini merupakan penjelasan *Source Code* dari Implementasi *filtering* yang ditunjukkan pada Tabel 4.28.

Tabel 4.28 Penjelasan *Source Code Filtering*

1-3	Memanggil *stopword list*
4	Memisahkan setiap string dengan *whitespace*
5	Menutup file yang terbaca

6 7	Menghapus nilai yang kosong pada *stopword list* `array_map ()` digunakan menghapus kata pada tokenisasi serta tersedia pada *stopword list*

4.11.1.5 Stemming

Pada tahap ini dilakukan pengubahan kata sehingga kata menjadi bentuk dasar kemudian dilanjutkan kembali pada tahap *filtering* sesuai dengan hasil *stemming*, hal ini dilakukan agar proses selanjutnya lebih optimal. Adapun algoritma *stemming* yang dimanfaatkan pada sistem ialah algoritma Nazief dan Andriani. Implementasi tahap *stemming* dapat dilihat pada *Source Code* 4.5.

```
1  $stemming = $filtering;
2  foreach ($stemming as &$daftarstemming){
3      $hapus1 = Del_Inflection_Suffixes($daftarstemming);
4      $hapus2 = Del_Derivation_Suffixes($hapus1);
5      $hapus3 = Del_Derivation_Prefix($hapus1);
6      $daftarstemming = $hapus3;
7      continue;
     }
```

Source Code* 4.5 Implementasi Tahap *Stemming

Berikut ini merupakan penjelasan *Source Code* dari Implementasi *stemming* yang ditunjukkan pada Tabel 4.29.

Tabel 4.29 Penjelasan *Source Code Stemming*

1-2	Memanggil data dari `$filtering` dilakukan perulangan dan pengecekan pada setiap kata
3-5	Menghapus *inflextion suffixes*, *derivation suffixes*, dan *derivation prefix*
6	Daftar kata *stemming* yang digunakan berada pada tahap terakhir, yakni sesudah melakukan penghapusan *derivation prefix*

4.11.2 Pembobotan *Term* (*Term Weighting*)

Untuk melakukan pembobotan *term* dilakukan perhitungn dengan menggunakan TF-IDF yang mana IDF menunjukan keunikan atu perbedaan pada kemunculan kata yang terdapat pada semua kumpulan dokumen. Persaman untuk melakukan pembobotan ialah persaman (2.1) hingga (2.4) yang mana proses perhitungan dilakukan hingga mendapatkan nilai $W_{t,d}$ yang sebelumnya sudah dinormalisasikan.

4.11.2.1 Implementasi *Term Weighting* ($W_{t,d}$)

Implementasi pembobotan *term* dapat dilihat pada *Source Code* 4.6 dimana dilakukan perhitungan bobot *term*.

```
1  $nilaibobot =0;
```

```
$hitungpembobotan = NULL;
    foreach ($listTF as $keyTF => $valueTF) {
      $nilaibobot = 1 + log10($valueTF);
      $hitungpembobotan[$keyTF]=$nilaibobot;
}
```

***Source Code* 4.6 Implementasi *Term Weighting* ($W_{t,d}$)**

Berikut ini merupakan penjelasan *Source Code* dari Implementasi *term weighting* ($W_{t,d}$) yang ditunjukkan pada Tabel 4.30.

Tabel 4.30 Penjelasan *Source Code Term Weighting* ($W_{t,d}$)

1-6	Melakukan pembobotan nilai data tf menggunakan persamaan untuk pembobotan *term*

4.11.2.2 Implementasi *Inverse Document Frequency* (IDF_t)

Implementasi *inverse document frequency* (IDF_t) dapat dilihat pada *Source Code* 4.7

```
foreach ($hitungpembobotan as $key => $value) {
      $listkata = $key;
      $takeIDF = "SELECT idf1 FROM normalisasi1 WHERE
kata='$listkata'";
      $selectIDF = mysqli_query($mysqli, $takeIDF);
      $row1=mysqli_fetch_array($selectIDF);
      $takeIDF = $row1[0];
      if($takeIDF != 0){
        $nilaiIDF[$listkata] = $takeIDF;
   }
}
```

***Source Code* 4.7 Implementasi *Inverse Document Frequency* (IDF_t)**

Berikut ini merupakan penjelasan *Source Code* dari Implementasi *inverse document frequency* (IDF_t) yang ditunjukkan pada Tabel 4.31.

Tabel 4.31 Penjelasan *Source Code Inverse Document Frequency* (IDF_t)

1-11	Mengambil nilai IDF_t pada *list* kata dalam *database*

4.11.2.3 Implementasi Perkalian TF dan IDF

Implementasi perkalian TF dan IDF dapat dilihat pada *Source Code* 4.8

```
$hitungTFIDF = NULL;
foreach ($nilaiIDF as $keyIDF => $valueIDF) {
      foreach ($hitungpembobotan as $key1 => $value1){
        if ($keyIDF == $key1){
          $hitungTFIDF[$key1] = $valueIDF * $value1;
} } }
```

***Source Code* 4.8 Implementasi Perkalian TF dan IDF**

Berikut ini merupakan penjelasan *Source Code* dari Implementasi *inverse document frequency* (IDF_t) yang ditunjukkan pada Tabel 4.32.

Tabel 4.32 Penjelasan *Source Code* Perkalian TF dan IDF

1-6	Melakukan perhitungan dengan mengkalikan nilai TF dengan IDF

4.11.2.4 Implementasi Normalisasi

Implementasi normalisasi dimulai dengan melakukan perhitungan akar pangkat untuk setiap dokumen, perhitungan akar pangkat dapat dilihat pada *Source Code* 4.9

```
1  $total = 0;
2  $hitungpangkat = 0;
3  foreach ($hitungTFIDF as $key => $value) {
4        $hitungpangkat = pow($value,2);
5        $total = $total + $hitungpangkat;
6  }
```

***Source Code* 4.9 Implementasi Normalisasi (1)**

Berikut ini merupakan penjelasan *Source Code* dari Implementasi normalisasi yang ditunjukkan pada Tabel 4.33.

Tabel 4.33 Penjelasan *Source Code* Normalisasi (1)

1-6	Melakukan perhitungan akar pangkat pada setiap dokumen

Selanjutnya dilakukan proses perhitungan normalisasi, perhitungan normalisasi dapat dilihat pada *Source Code* 4.10

```
1  $hasilnormalisasi = NULL;
2  foreach ($hitungTFIDF as $key => $value) {
3        $hitungnormalisasi = $value/$total;
4        $hasilnormalisasi[$key]=$hitungnormalisasi;
5  }
```

***Source Code* 4.10 Implementasi Normalisasi (2)**

Berikut ini merupakan penjelasan *Source Code* dari Implementasi normalisasi yang ditunjukkan pada Tabel 4.34.

Tabel 4.34 Penjelasan *Source Code* Normalisasi (2)

1-6	Melakukan perhitungan normalisasi untuk setiap dokumen

4.11.3 Klasifikasi Improved K-NN

Setelah melakukan pembobotan *term* dan mendapatkan hasil normalisasi maka proses yang akan dilakukan selanjutnya adalah melakukan perhitungan pada nilai *cosine similarity* setelah mendapatkan hasil perhitungan dari *cosine similarity* maka tahap selanjutnya akan dilakukan pengklasifikasian sesuai dengan hasil dari *cosine similarity*. Untuk melakukan pernghitungan pada nilai *cosine similarity* dapat menggunakan persamaan (2.5) dimana perhitungan dapat dilakukan setelah melakukan normalisasi pada nilai $W_{t,d}$ terlebih dahulu, hal ini dilakukan agar nilai $W_{t,d}$ lebih terstruktur dan perhitungan *cosine similarity* akan lebih mudah dilakukan.

4.11.3.1 Perhitungan *Cosine Similarity*

Perhitungan nilai *cosine similarity* yang dilakukan ialah data latih terhadap data uji yang kategorinya ingin diketahui. *Source Code* 4.11 ialah implementasi dari perhitungan nilai *cosine similarity* yang mana menggunakan data latih sejumlah 250 dokumen serta hasil dari perhitungan *cosine similarity* langsung diurutkan berdasarkan tingkat kemiripan dokumen data latih terhadap data uji.

```
1   for ($i = 1; $i <= 20; $i++){
2     $cossimresult = 0;
3     $dl = "dl".$i;
4      foreach ($hasilnormalisasi1 as $key => $value) {
5        $takenormalisasi = "SELECT $dl FROM normalisasi1 WHERE
    kata='$key'";
6        $selectnormalisasi = mysqli_query($mysqli,
    $takenormalisasi);
7         $row1=mysqli_fetch_array($selectnormalisasi);
8       $hasil = $row1[0];
9       if($hasil != 0){
10        $hitungcossim = $hasil * $value;
11        $cossimresult = $cossimresult + $hitungcossim;
12      }
13      else{
14        $hitungcossim = 0;
15        $cossimresult = $cossimresult + $hitungcossim;
16      }
17    }
18    if($cossimresult != 0){
19       $arraysinopsis[$dl] = $cossimresult;
20      }
21    }
```

Source Code* 4.11 Implementasi *Cosine Similarity

Berikut ini merupakan penjelasan *Source Code* dari Implementasi *cosine similarity* yang ditunjukkan pada Tabel 4.35.

Tabel 4.35 Penjelasan *Source Code Cosine Similarity*

1-7	Melakukan pengecekan pada setiap kata serta mengambil data dari *database*, yakni data dari data nomalisasi
8-21	Proses perhitungan *cosine similarity*

4.11.3.2 Klasifikasi dengan Metode Improved K-NN

Proses perhitungan untuk melakukan klasifikasi dengan menggunakan metode Improved K-NN dimulai dengan menetapkan *k-values* awal, lalu menghitung *k-values* baru (n) yang ditunjukan oleh persamaan (2.6). Pada tahap pengujian, nilai *k-values* awal yang ditetapkan ialah 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 75, dan 100. Pada *source code* 5.8 ditunjukan implementasi yang mana nilai *k-values* awal memliki nilai 2 dengan data latih sebanyak sebanyak 100 dokumen. Implementasi ditunjukan pada *Source Code* 4.12.

```
1   $nilaik = [2,4,6,8,10,15,20,25,30,35,40,45,50,75,100];
2     foreach ($nilaik as $keyk => $valuek) {
3       $kvalues = $valuek;
```

```
4       $kRomantis = ($kvalues*50)/50;
5       $kRomantis = round($kRomantis);
6       $hitung = 0;
7       $jumlahRomantis = 0;

8       foreach ($arraysinopsis as $key => $value) {
9         $hitung = $hitung+1;
10        if($hitung<=$kRomantis){
11          for ($j = 1; $j <= 50; $j++){
12            $dl = "dl".$j;
13            if ($key==$dl){
14              $kategori = "Romantis";
15            }
16          }
17          if($kategori=="Romantis"){
18            $romantis = $romantis + $value;
19          }
20          else{
21            $romantis = $romantis + 0;
22          }
23        }
24      }
25      $probRomantis = $romantis/$jumlahRomantis;
26      if($probRomantis > $probKeluarga){
27            if($probRomantis > $probAksi){
28              if($probRomantis > $probHoror){
29                if($probRomantis > $prob_Thriller_){
30                  $kategoriakhir = "Romantis";
31                }
32                elseif ($probRomantis < $prob_Thriller_){
33                  $kategoriakhir = "_Thriller_";
34                }
35                else{
36                  $kategoriakhir ="Seri untuk beberapa
37      Kategori";
38                }
39              }
40              elseif ($probRomantis < $probHoror){
41                $kategoriakhir = "Horor";
42              }
43              else{
44                $kategoriakhir ="Seri untuk beberapa
45      Kategori";
46              }
47            }
48            elseif ($probRomantis < $probAksi){
49              $kategoriakhir = "Aksi";
50            }
51            else{
52              $kategoriakhir ="Seri untuk beberapa Kategori";
53            }
54          }
55      }
56      echo "k-Values = ".$kvalues." (".$kategoriakhir.")
57      <br>";
58      $kategoriawal = $kategoriawal;
59      $kategoriakhir = $kategoriakhir;
```

***Source Code* 4.12 Implementasi Improved K-NN**

Berikut ini merupakan penjelasan *Source Code* dari Implementasi Improved K-NN yang ditunjukkan pada Tabel 4.36.

Tabel 4.36 Penjelasan *Source Code* Improved K-NN

1	Nilai *k* awal yakni 2,4,6,8,10,15,20,25,30,35,40,45,50,75, dan 100
2-24	Menghitung nilai *k* baru (*n*) dengan data latih sebanyak 50
25-59	Melakukan klasifikasi dengan algoritma Improved K-NN, dimana sebelumnya menghitung probabilitas kategori terlebih dahulu, kemudian dilakukan perbandingan pada setiap kategori, peluang kategori terbesar akan menjadi acuan hasil kategori.

4.12 Implementasi Antar Muka

Antarmuka (*interface*) sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN di bangun demi memberikan kemudahan bagi pengguna untuk melakukan interaksi dengan sistem.

4.12.1 Tampilan Halaman Awal

Pada halaman awal terdapat judul dari sistem dan dua buah tombol utama, yakni tombol pengujian yang memiliki fungsi untuk melakukan akses pada halaman pengujian dan tombol halaman pengguna yang berfungsi untuk mengalihkan pengguna ke halaman pengguna. Gambar halaman awal ditunjukan pada Gambar 4.12.

Gambar 4.15 Tampilan Halaman Awal

4.12.2 Tampilan Halaman Pengujian

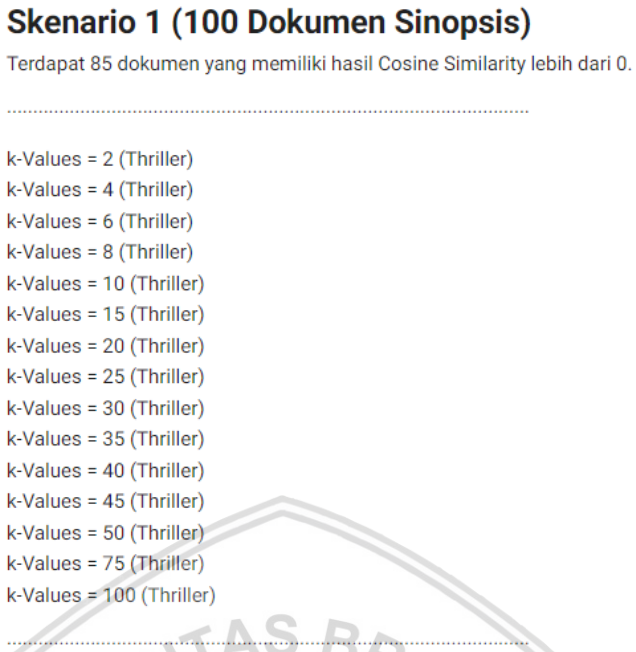
Pada halaman pengujian terdapat judul dari sistem serta dua buah kolom untuk memasukan data berupa sinopsis film dan memilih kategori awal untuk data yang dimasukan, selain itu terdapat pula dua buah tombol yang memiliki fungsi untuk memproses dokumen dan untuk kembali ke halaman awal. Gambar halaman pengujian ditunjukan pada Gambar 4.13.

Gambar 4.16 Tampilan Halaman Pengujian

4.12.3 Tampilan Halaman Hasil Pengujian

Pada halaman hasil pengujian menunjukan hasil dari pengklasifikasian terhadap data uji yang telah dimasukan sebelumnya, dimana data uji terlebih dahulu telah melewati beberapa proses yang dimulai dari proses *preprocessing*, pembobotan *term,* hingga melakukan klasifikasi dengan menggunakan metode Improved K-NN. Gambar halaman hasil pengujian ditunjukan oleh Gambar 4.14 dan Gambar 4.15.

Gambar 4.17 Tampilan Halaman Hasil Pengujian (1)



Gambar 4.18 Tampilan Halaman Hasil Pengujian (2)

4.12.4 Tampilan Halaman Pengguna

Pada halaman pengguna terdapat judul sistem dan kolom yang digunakan untuk memasukan sinopsis film, selain itu terdapat pula dua buah tombol yang mana memiliki fungsi untuk meproses data yang berupa sinopsis tersebut dan tombol untuk kembali kehalaman awal. Gambar halaman pengguna ditunjukan oleh Gambar 4.16.

Gambar 4.19 Tampilan Halaman Pengguna

4.12.5 Tampilan Halaman Hasil Pengujian Pengguna

Pada halaman hasil pengujian pengguna ialah halaman yang menunjukkan hasil kategori dari data yang di masukan oleh pengguna. Gambar halaman hasil pengujian pengguna ditunjukan pada Gambar 4.17 hingga Gambar 4.22.

Klasifikasi Film Berdasarkan Sinopsis Menggunakan Improved K-NN

Hasil Pengujian

Sinopsis : bercerita tentang sersan David Tubbs (Brandon Smith) mencoba membantu tim penyidik untuk menghentikan Creeper agar tidak membunuh orang-orang tak bersalah dengan mempelajari rahasia asal-usulnya yang sesungguhnya. Plot ceritanya berkisah sekelompok pemburu yang sangat menyadari akan eksistensi sang Creepers dan berniat untuk memburu sang mahluk jejadian agar tidak menebar teror lagi.

HASIL
Cleansing : bercerita tentang sersan David Tubbs Brandon Smith mencoba membantu tim penyidik untuk menghentikan Creeper agar tidak membunuh orang orang tak bersalah dengan mempelajari rahasia asal usulnya yang sesungguhnya Plot ceritanya berkisah sekelompok pemburu yang sangat menyadari akan eksistensi sang Creepers dan berniat untuk memburu sang mahluk jejadian agar tidak menebar teror lagi

Case Folding : bercerita tentang sersan david tubbs brandon smith mencoba membantu tim penyidik untuk menghentikan creeper agar tidak membunuh orang orang tak bersalah dengan mempelajari rahasia asal usulnya yang sesungguhnya plot ceritanya berkisah sekelompok pemburu yang sangat menyadari akan eksistensi sang creepers dan berniat untuk memburu sang mahluk jejadian agar tidak menebar teror lagi

Tokenisasi : agar | agar | akan | asal | bercerita | berkisah | berniat | bersalah | brandon | ceritanya | creeper | creepers | dan | david | dengan | eksistensi | jejadian | lagi | mahluk | membantu | membunuh | memburu | mempelajari | mencoba | menebar | menghentikan | menyadari | orang | orang | pemburu | penyidik | plot | rahasia | sang | sang | sangat | sekelompok | sersan | sesungguhnya | smith | tak | tentang | teror | tidak | tidak | tim | tubbs | untuk | untuk | usulnya | yang | yang

Gambar 4.20 Tampilan Halaman Hasil Pengujian Pengguna (1)

Stemming : cerita | kisah | niat | sa | brandon | cerita | creeper | creepers | david | eksistensi | jejadian | mahluk | bantu | bunuh | buru | mempelajari | coba | tebar | henti | sadar | orang | orang | buru | sidik | plot | rahasia | sang | sang | kelompok | sersan | sungguh | smith | teror | tim | tubbs | usul

.....

TF
cerita : 2
kisah : 1
niat : 1
sa : 1
brandon : 1
creeper : 1
creepers : 1
david : 1
eksistensi : 1
jejadian : 1
mahluk : 1
bantu : 1
bunuh : 1
buru : 2
mempelajari : 1
coba : 1
tebar : 1
henti : 1
sadar : 1
orang : 2
sidik : 1
plot : 1
rahasia : 1

Gambar 4.21 Tampilan Halaman Hasil Pengujian Pengguna (2)

Pembobotan TF
cerita : 1.301029995664
kisah : 1
niat : 1
sa : 1
brandon : 1
creeper : 1
creepers : 1
david : 1
eksistensi : 1
jejadian : 1
mahluk : 1
bantu : 1
bunuh : 1
buru : 1.301029995664
mempelajari : 1
coba : 1
tebar : 1
henti : 1
sadar : 1
orang : 1.301029995664
sidik : 1
plot : 1
rahasia : 1
sang : 1.301029995664
kelompok : 1
sersan : 1
sungguh : 1
smith : 1
teror : 1
tim : 1
tubbs : 1
usul : 1

Gambar 4.22 Tampilan Halaman Hasil Pengujian Pengguna (3)

Nilai IDF
cerita : 0.0861861
kisah : 0.29243
niat : 1.52288
sa : 1.82391
david : 2.30103
eksistensi : 2.30103
bantu : 1.04576
bunuh : 0.619789
buru : 1.25964
coba : 1.04576
tebar : 2.30103
henti : 1.1549
sadar : 1.04576
orang : 0.59346
sidik : 2.30103
rahasia : 1.18709
sang : 0.732828
kelompok : 0.886057
sersan : 2
sungguh : 2
smith : 2.30103
teror : 1.12494
tim : 1.39794
usul : 2

Gambar 4.23 Tampilan Halaman Hasil Pengujian Pengguna (4)

Perkalian TF-IDF
cerita : 0.1121307013093
kisah : 0.29243
niat : 1.52288
sa : 1.82391
david : 2.30103
eksistensi : 2.30103
bantu : 1.04576
bunuh : 0.619789
buru : 1.6388294237382
coba : 1.04576
tebar : 2.30103
henti : 1.1549
sadar : 1.04576
orang : 0.77210926122675
sidik : 2.30103
rahasia : 1.18709
sang : 0.95343120966244
kelompok : 0.886057
sersan : 2
sungguh : 2
smith : 2.30103
teror : 1.12494
tim : 1.39794
usul : 2

.....

Gambar 4.24 Tampilan Halaman Hasil Pengujian Pengguna (5)

Normalisasi
cerita : 0.014620329199191
kisah : 0.038128922924742
niat : 0.19856298650491
sa : 0.23781323329229
david : 0.30002323809978
eksistensi : 0.30002323809978
bantu : 0.13635298169742
bunuh : 0.080812115756258
buru : 0.21368122553948
coba : 0.13635298169742
tebar : 0.30002323809978
henti : 0.15058336383334
sadar : 0.13635298169742
orang : 0.10067262083505
sidik : 0.30002323809978
rahasia : 0.15478050512851
sang : 0.12431455427713
kelompok : 0.11552986718164
sersan : 0.26077299131239
sungguh : 0.26077299131239
smith : 0.30002323809978
teror : 0.14667698442348
tim : 0.18227249773762
usul : 0.26077299131239

Berdasarkan Skenario 5 (200 Data Latih dan K-Values = 40)

Masuk ke dalam Kategori Thriller.

Gambar 4.25 Tampilan Halaman Hasil Pengujian Pengguna (6)

BAB 5 PENGUJIAN DAN ANALISIS

5.1 Pengujian

5.1.1 *Precission, Recall, F-Measure* dan Akurasi

Agar dapat mengetahui pengaruh serta jumlah data latih dan juga *k-values* pada keberhasilan sistem klasifikasi, maka dilakukan proses pengujian dengan menggunakan beberapa skenario. Masing-masing skenario memiliki jumlah data latih serta *k-values* awal yang berbeda-beda yang mana semua skenario menggunakan jumlah data uji yang sama yaitu sebanyak 50 dokumen yang menjadi data uji. Pada Tabel 5.1 menunjukan skenario yang dibangun.

Tabel 5.1 Skenario Pengujian

Skenario	Data Latih						Data Uji					
	K1	K2	K3	K4	K5	Jumlah	K1	K2	K3	K4	K5	Jumlah
1	20	25	15	10	30	100	10	10	10	10	10	50
2	25	30	35	40	20	150	10	10	10	10	10	50
3	35	20	50	40	30	175	10	10	10	10	10	50
4	40	35	45	30	50	200	10	10	10	10	10	50
5	50	25	40	40	45	200	10	10	10	10	10	50

Keterangan :

- K1 = Romantis
- K2 = Keluarga
- K3 = Aksi
- K4 = Horor
- K5 = *Thriller*

5.1.2 Skenario 1

Pengujian pada skenario 1 dilakukan dengan menggunakan data latih sejumlah 100 dokumen dimana pada setiap kategori memiliki jumlah data latih yang berbeda. Pada kategori romantis memiliki data latih sebanyak 20, untuk kategori keluarga memiliki data latih sebanyak 25, untuk kategori aksi memiliki data latih sebanyak 15, untuk kategori horor memiliki data latih sebanyak 10, dan untuk kategori *thriller* memiliki data latih sebanyak 20, pengujian pada skenario 1 dilakukan dengan menggunakan 50 data uji dimana setiap kategori memiliki jumlah data uji yang sama yakni masing-masing menggunakan 10 data uji .

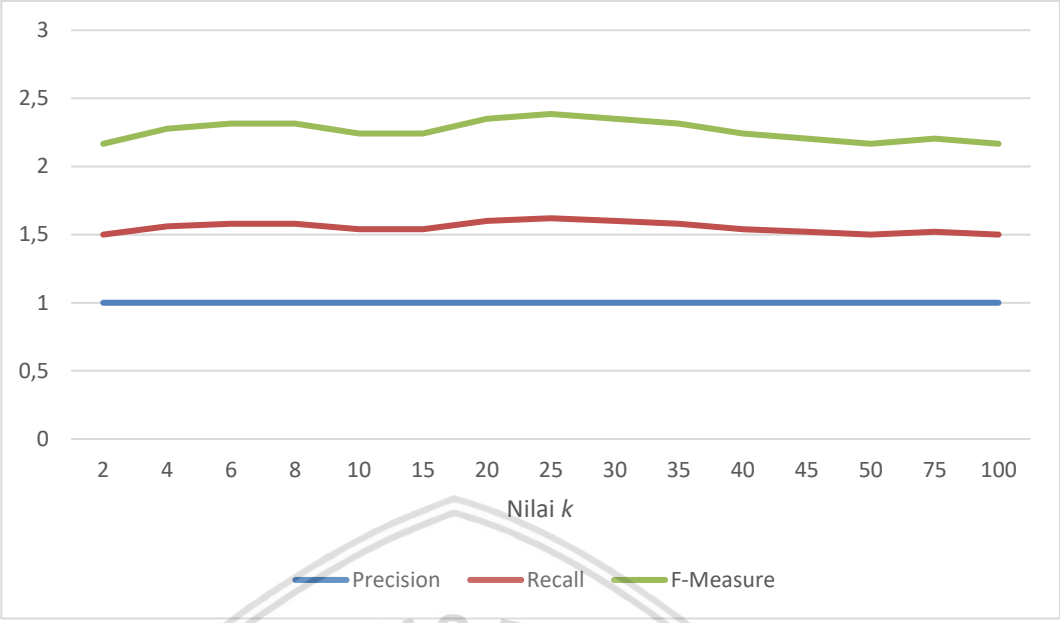
Tabel 5.2 *Precision, Recall, F-Measure,* dan Akurasi pada Skenario 1

k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	0	0	0	0	0	1	0,5	0,66666667	50%
4	1	2	1	1	2	1	0,56	0,71794872	56%
6	3	3	2	1	4	1	0,58	0,73417722	58%
8	4	5	3	2	6	1	0,58	0,73417722	58%
10	5	7	4	3	8	1	0,54	0,7012987	54%
15	7	8	5	3	10	1	0,54	0,7012987	54%
20	10	13	8	5	15	1	0,6	0,75	60%
25	13	17	10	7	20	1	0,62	0,7654321	62%
30	17	21	13	8	25	1	0,6	0,75	60%
35	20	25	15	10	30	1	0,58	0,73417722	58%
40	23	29	18	12	35	1	0,54	0,7012987	54%
45	27	33	20	13	40	1	0,52	0,68421053	52%
50	30	38	23	15	45	1	0,5	0,66666667	50%
75	33	42	25	17	50	1	0,52	0,68421053	52%
100	50	63	38	25	75	1	0,5	0,66666667	50%

Keterangan :

- K1 = Romantis
- K2 = Keluarga
- K3 = Aksi
- K4 = Horor
- K5 = *Thriller*

Pada Tabel 5.2 menunjukan hasil dari masing-masing nilai *precision, recall, f-measure* dan akurasi dari pengujian dengan menggunakan skenario 1. Dimana dilakukan perhitungan kembali pada *k-values* awal sehingga kemudian didapatkan nilai *n* atau *k-values* baru untuk setiap kategori yang dihitung dengan menggunakan persamaan (2.6). Skenario 1 menunjukan dimana, nilai *f-measure* yang tertinggi berada pada *k-values* awal dengan nilai 20 dan 25 yang memiliki nilai sebesar 0,7654321 dan nilai *f-measure* yang terendah berada pada *k-values* awal dengan nilai 2, 50 dan 100 yang memiliki nilai sebesar 0,666666667.



Gambar 5.1 Grafik Hasil Pengujian dengan Skenario 1

Gambar 5.1 memperlihatkan grafik dari hasil pengujian dengan skenario 1 dimana garis berwarna biru mewakili nilai *precision*, garis berwarna merah mewakili nilai *recall*, dan garis warna hijau mewakili nilai *f-measure.* Garis *f-measure* menunjukan peningkatan pada nilai k = 25 sedangkan pada nilai k = 2 garis menurun. Hal ini menunjukan bahwa nilai *f-measure* tertinggi berada pada nilai k = 20 dan terendah berada pada nilai k = 2.

5.1.3 Skenario 2

Pengujian pada skenario 2 dilakukan dengan menggunakan data latih sejumlah 150 dokumen yang mana untuk kategori romantis berjumlah 25, untuk kategori keluarga berjumlah 30, untuk kategori aksi berjumlah 35, untuk kategori horor berjumlah 40, dan untuk kategori *thriller* berjumlah 20, dengan menggunakan 50 data uji.

Tabel 5.3 *Precision, Recall, F-Measure,* dan Akurasi pada Skenario 2

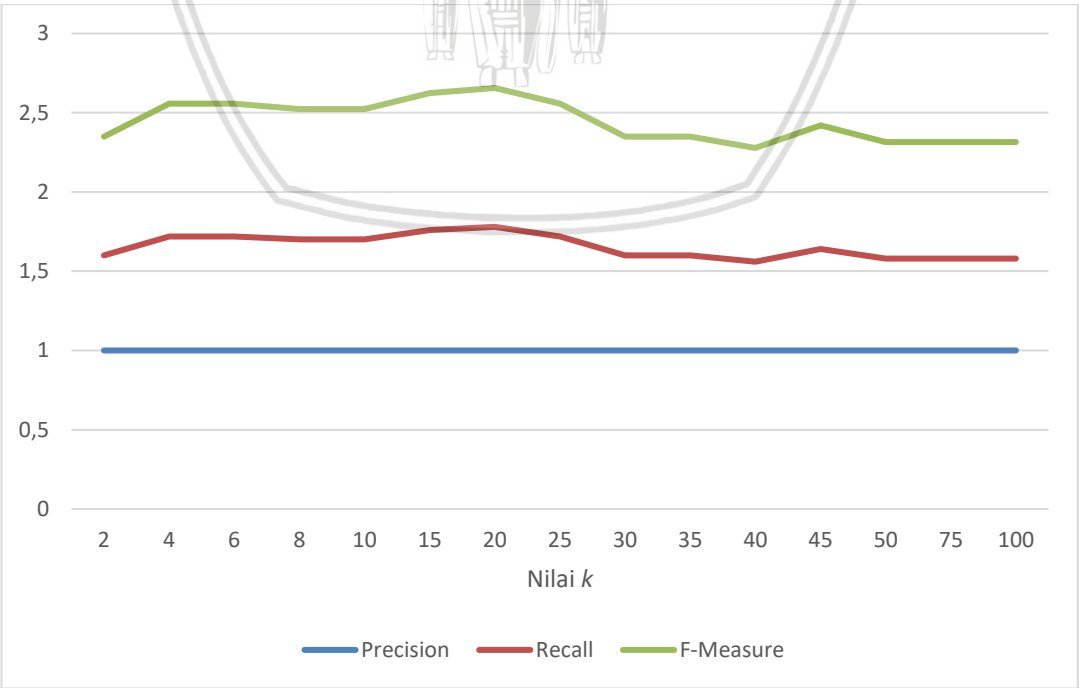
k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	1	2	2	2	1	1	0,6	0,75	60%
4	3	3	4	4	2	1	0,72	0,8372093	72%
6	4	5	5	6	3	1	0,72	0,8372093	72%
8	5	6	7	8	4	1	0,7	0,82352941	70%
10	6	8	9	10	5	1	0,7	0,82352941	70%
15	9	11	13	15	8	1	0,76	0,86363636	76%
20	13	15	18	20	10	1	0,78	0,87640449	78%

25	16	19	22	25	13	1	0,72	0,8372093	72%
30	19	23	26	30	15	1	0,6	0,75	60%
35	22	26	31	35	18	1	0,6	0,75	60%
40	25	30	35	40	20	1	0,56	0,71794872	56%
45	28	34	39	45	23	1	0,64	0,7804878	64%
50	31	38	44	50	25	1	0,58	0,73417722	58%
75	47	56	66	75	38	1	0,58	0,73417722	58%
100	63	75	88	100	50	1	0,58	0,73417722	58%

Keterangan :

- K1 = Romantis
- K2 = Keluarga
- K3 = Aksi
- K4 = Horor
- K5 = *Thriller*

Tabel 5.3 menunjukan hasil dari masing-masing nilai *precision, recall, f-measure* dan akurasi dari pengujian dengan menggunakan skenario 2. Dimana dilakukan perhitungan kembali pada *k-values* awal sehingga mendapatkan nilai n atau *k-values* baru untuk setiap kategori yang dihitung menggunakan persamaan (2.6). Skenario 2 menunjukan, nilai *f-measure* yang tertinggi berada pada *k-values* awal dengan nilai 20 yakni sebesar 0,876404494 dan nilai *f-measure* yang terendah berada pada *k-values* awal dengan nilai 40 yakni sebesar 0,717948718.



Gambar 5.2 Grafik Hasil Pengujian dengan Skenario 2

Gambar 5.2 memperlihatkan grafik dari hasil pengujian dengan skenario 2 dimana garis berwarna biru mewakili nilai *precision*, garis berwarna merah mewakili nilai *recall*, dan garis warna hijau mewakili nilai *f-measure.* Garis *f-measure* menunjukan peningkatan pada nilai *k* = 20 sedangkan pada nilai *k* = 40 garis mengalami penurunan. Hal ini menunjukan bahwa nilai *f-measure* tertinggi berada pada nilai *k* = 20 dan terendah berada pada nilai *k* = 40.

5.1.4 Skenario 3

Pengujian pada skenario 3 dilakukan dengan menggunakan data latih sejumlah 175 dokumen yang mana untuk kategori romantis memiliki data latih yang berjumlah 35, untuk kategori keluarga memiliki data latih berjumlah 20, untuk kategori aksi memiliki data latih berjumlah 50, untuk kategori horor memiliki data latih berjumlah 40, dan untuk kategori *thriller* memiliki data latih berjumlah 30, dengan menggunakan 50 data uji yang untuk setiap kategorinya memiliki jumlah yang sama yakni sebesar 10 data uji.

Tabel 5.4 *Precision, Recall, F-Measure,* dan Akurasi pada Skenario 3

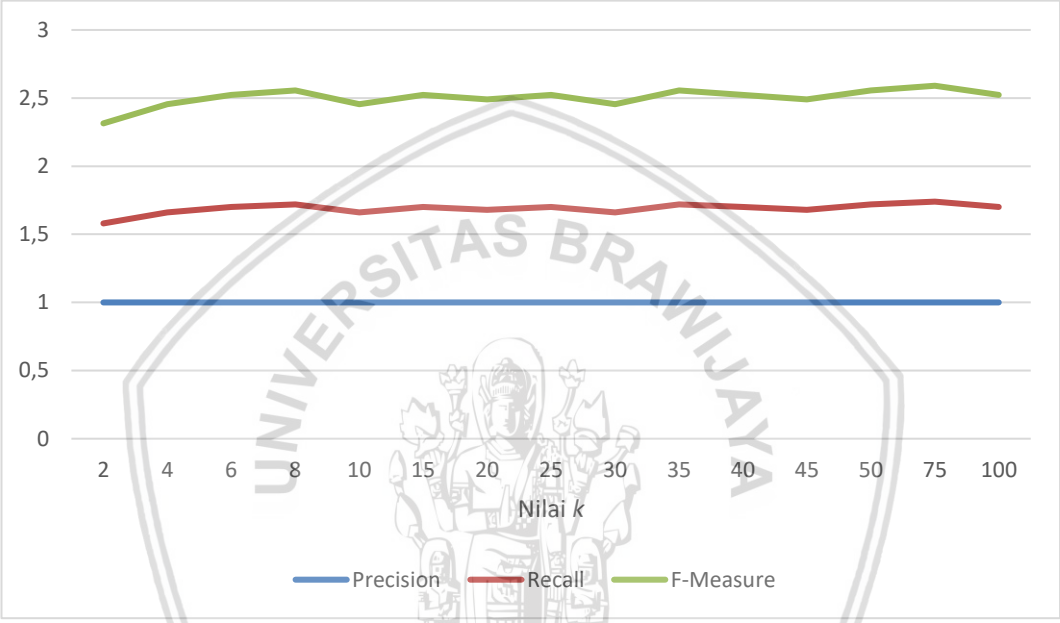
k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	1	1	2	2	1	1	0,58	0,73417722	58%
4	3	2	4	3	2	1	0,66	0,79518072	66%
6	4	2	6	5	4	1	0,7	0,82352941	70%
8	6	3	8	6	5	1	0,72	0,8372093	72%
10	7	4	10	8	6	1	0,66	0,79518072	66%
15	11	6	15	12	9	1	0,7	0,82352941	70%
20	14	8	20	16	12	1	0,68	0,80952381	68%
25	18	10	25	20	15	1	0,7	0,82352941	70%
30	21	12	30	24	18	1	0,66	0,79518072	66%
35	25	14	35	28	21	1	0,72	0,8372093	72%
40	28	16	40	32	24	1	0,7	0,82352941	70%
45	32	18	45	36	27	1	0,68	0,80952381	68%
50	35	20	50	40	30	1	0,72	0,8372093	72%
75	53	30	75	60	45	1	0,74	0,85057471	74%
100	70	40	100	80	60	1	0,7	0,82352941	70%

Keterangan :

- K1 = Romantis
- K2 = Keluarga
- K3 = Aksi

- K4 = Horor
- K5 = *Thriller*

Tabel 5.4 menunjukan hasil dari masing-masing nilai *precision, recall, f-measure* dan akurasi dari pengujian dengan menggunakan skenario 3. Dilakukan perhitungan kembali dari *k-values* awal sehingga menjadi n (*k-values* baru) untuk setiap kategori yang dihitung menggunakan persamaan (2.6). Skenario 3 menunjukan, nilai *f-measure* tertinggi ada pada *k-values* awal dengan nilai 75 yakni sebesar 0,850574713 dan terendah ada pada *k-values* awal dengan nilai 2 yakni sebesar 0,734177215.



Gambar 5.3 Grafik Hasil Pengujian dengan Skenario 3

Gambar 5.3 memperlihatkan grafik dari hasil pengujian dengan skenario 3 dimana garis berwarna biru mewakili nilai *precision*, garis berwarna merah mewakili nilai *recall*, dan garis warna hijau mewakili nilai *f-measure.* Garis *f-measure* menunjukan bahwa terjadi peningkatan pada nilai *k* = 75 sedangkan pada nilai *k* = 2 garis menujukan bahwa terjadinya penurunan. Hal ini menunjukan bahwa nilai *f-measure* tertinggi berada pada nilai *k* = 75 dan nilai *f-measure* terendah berada pada nilai *k* = 2.

5.1.5 Skenario 4

Pengujian pada skenario 4 dilakukan dengan menggunakan data latih sejumlah 200 dokumen yang mana untuk kategori romantis memiliki data latih berjumlah 50, untuk kategori keluarga memiliki data latih berjumlah 25, untuk kategori aksi memiliki data latih berjumlah 40, untuk kategori horor memiliki data latih berjumlah 40, dan untuk kategori *thriller* memiliki data latih berjumlah 45 , dengan menggunakan 50 data uji yang untuk setiap kategorinya memiliki jumlah yang sama yakni sebesar 10 data uji.

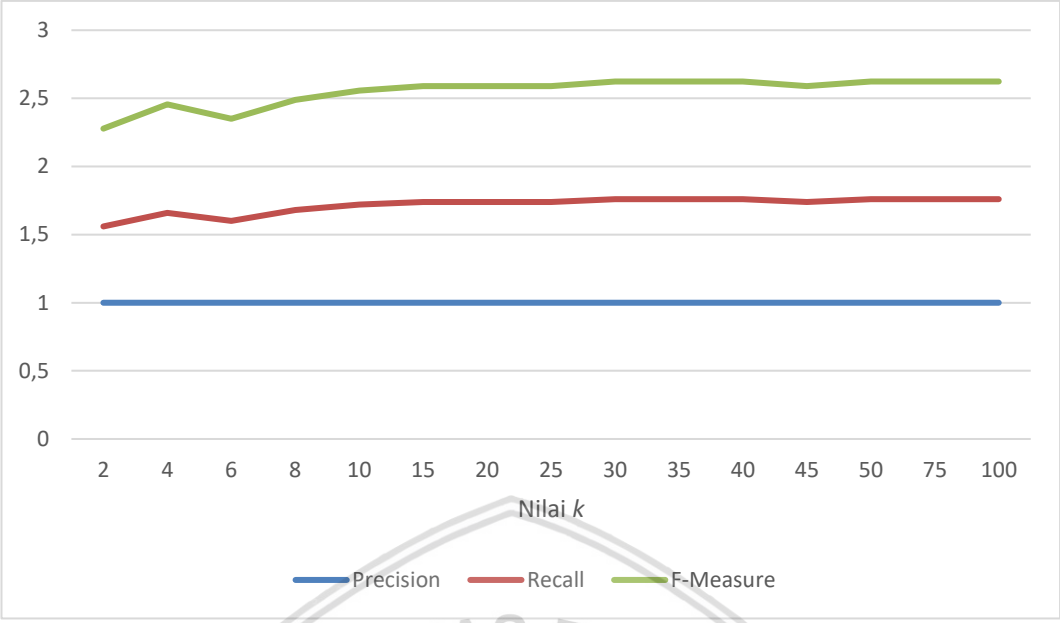
Tabel 5.5 *Precision, Recall, F-Measure,* dan Akurasi pada Skenario 4

k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	2	1	2	2	2	1	0,56	0,71794872	56%
4	4	2	3	3	4	1	0,66	0,79518072	66%
6	6	3	5	5	5	1	0,6	0,75	60%
8	8	4	6	6	7	1	0,68	0,80952381	68%
10	10	5	8	8	9	1	0,72	0,8372093	72%
15	15	8	12	12	14	1	0,74	0,85057471	74%
20	20	10	16	16	18	1	0,74	0,85057471	74%
25	25	13	20	20	23	1	0,74	0,85057471	74%
30	30	15	24	24	27	1	0,76	0,86363636	76%
35	35	18	28	28	32	1	0,76	0,86363636	76%
40	40	20	32	32	36	1	0,76	0,86363636	76%
45	45	23	36	36	41	1	0,74	0,85057471	74%
50	50	25	40	40	45	1	0,76	0,86363636	76%
75	75	38	60	60	68	1	0,76	0,86363636	76%
100	100	50	80	80	90	1	0,76	0,86363636	76%

Keterangan :

- K1 = Romantis
- K2 = Keluarga
- K3 = Aksi
- K4 = Horor
- K5 = *Thriller*

Tabel 5.5 menunjukan hasil dari masing-masing nilai *precision, recall, f-measure* dan akurasi dari pengujian dengan menggunakan skenario 4. Dimana dilakukan perhitungan kembali pada *k-values* awal sehingga didapatkan nilai *n* atau *k-values* baru untuk setiap kategori yang dihitung dengan menggunakan persamaan (2.6). Skenario 4 menunjukan hasil berupa, nilai *f-measure* tertinggi berada pada *k-values* awal dengan nilai 30, 35, 40, 50, 75 dan 100 dengan nilai sebesar 0,863636364 dan nilai *f-measure* terendah berada pada *k-values* awal dengan nilai 2 dengan nilai sebesar 0,717948718.



Gambar 5.4 Grafik Hasil Pengujian dengan Skenario 4

Gambar 5.4 memperlihatkan grafik dari hasil pengujian dengan skenario 4 dimana garis berwarna biru mewakili nilai *precision*, garis berwarna merah mewakili nilai *recall*, dan garis warna hijau mewakili nilai *f-measure.* Garis *f-measure* menunjukan peningkatan pada nilai *k* = 40, 50, 75, dan 100 sedangkan pada nilai *k* = 2 garis mengalami penurunan. Hal ini menunjukan bahwa nilai *f-measure* tertinggi berada pada nilai *k* = 40, 50, 75 dan 100 dan terendah berada pada nilai *k* = 2.

5.1.6 Skenario 5

Pengujian pada skenario 5 dilakukan dengan menggunakan data latih sejumlah 200 dokumen yang mana untuk kategori romantis berjumlah 40, untuk kategori keluarga berjumlah 35, untuk kategori aksi berjumlah 45, untuk kategori horor berjumlah 30, dan untuk kategori *thriller* berjumlah 50, dengan menggunakan 50 data uji.

Tabel 5.6 *Precision, Recall, F-Measure,* dan Akurasi pada Skenario 5

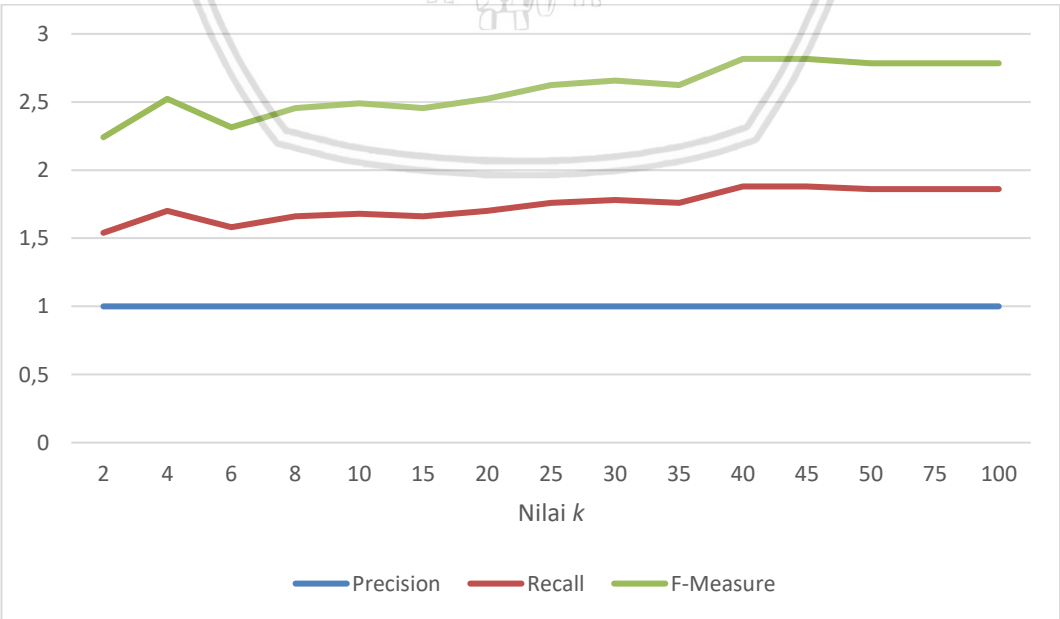
k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	2	1	2	1	2	1	0,54	0,701298701	54%
4	3	3	4	2	4	1	0,7	0,823529412	70%
6	5	4	5	4	6	1	0,58	0,734177215	58%
8	6	6	7	5	8	1	0,66	0,795180723	66%
10	8	7	9	6	10	1	0,68	0,80952381	68%
15	12	11	14	9	15	1	0,66	0,795180723	66%

20	16	14	18	12	20	1	0,7	0,823529412	70%
25	20	18	23	15	25	1	0,76	0,863636364	76%
30	24	21	27	18	30	1	0,78	0,876404494	78%
35	28	25	32	21	35	1	0,76	0,863636364	76%
40	32	28	36	24	40	1	0,88	0,936170213	88%
45	36	32	41	27	45	1	0,88	0,936170213	88%
50	40	35	45	30	50	1	0,86	0,924731183	86%
75	60	53	68	45	75	1	0,86	0,924731183	86%
100	80	70	90	60	100	1	0,86	0,924731183	86%

Keterangan :

- K1 = Romantis
- K2 = Keluarga
- K3 = Aksi
- K4 = Horor
- K5 = *Thriller*

Tabel 5.7 menunjukan hasil dari masing-masing nilai *precision, recall, f-measure* dan akurasi dari pengujian dengan menggunakan skenario 5. Dilakukan perhitungan kembali dari *k-values* awal sehingga menjadi n (*k-values* baru) untuk setiap kategori yang dihitung menggunakan persamaan (2.6). Skenario 5 menunjukan, nilai *f-measure* tertinggi ada pada *k-values* awal dengan nilai 40 dan 45 yakni sebesar 0,936170213 dan terendah ada pada *k-values* awal dengan nilai 2 yakni sebesar 0,7012987.



Gambar 5.5 Grafik Hasil Pengujian dengan Skenario 5

Gambar 5.5 memperlihatkan grafik dari hasil pengujian dengan skenario 5 dimana garis berwarna biru mewakili nilai *precision*, garis berwarna merah mewakili nilai *recall*, dan garis warna hijau mewakili nilai *f-measure.* Garis *f-measure* menunjukan peningkatan pada nilai *k* = 40, dan 45 sedangkan pada nilai *k* = 2 garis mengalami penurunan. Hal ini menunjukan bahwa nilai *f-measure* tertinggi berada pada nilai *k* = 40 dan 45 dan terendah berada pada nilai *k* = 2.

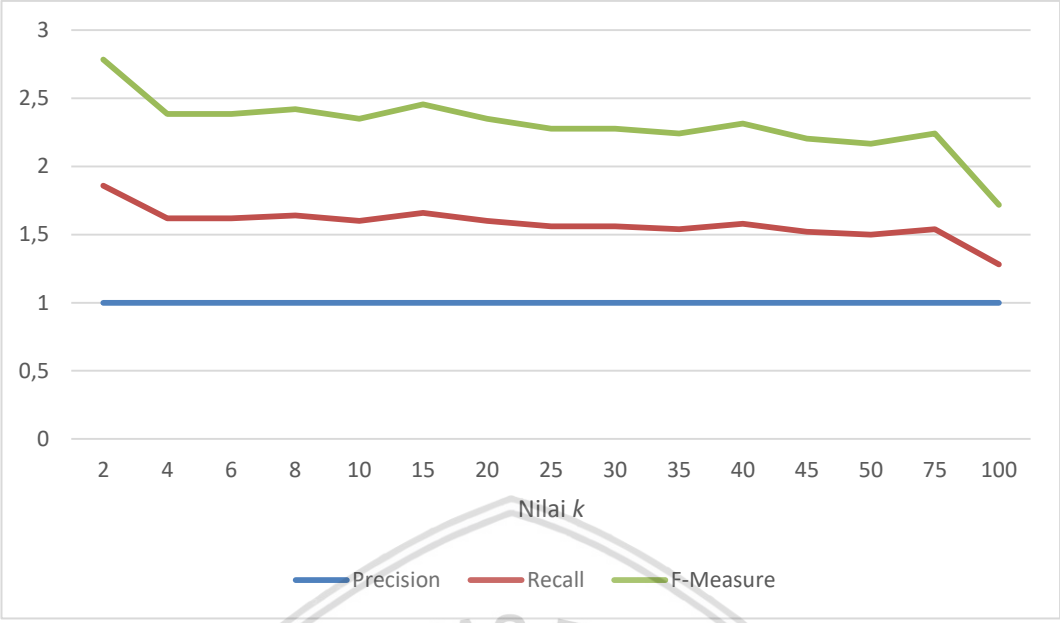
5.1.7 Perbandingan Hasil K-NN

Data latih pada skenario 5 digunakan untuk melakukan perbandingan hasil pengujian dengan metode K-NN pada skenario pengujian dengan metode K-NN. Skenario 5 dipilih karena hasil pengujiannya memiliki hasil yang terbaik dibandingkan dengan skenario lainnya yakni menggunakan 200 dokumen yang mana untuk kategori romantis berjumlah 40, untuk kategori keluarga berjumlah 35, untuk kategori aksi berjumlah 45, untuk kategori horor berjumlah 30, dan untuk kategori *thriller* berjumlah 50, dengan menggunakan 50 data uji. Tabel 5.8 ditunjukan hasil nilai dari *precision*, *recall*, *f-measure* dan akurasi yang dihasilkan.

Tabel 5.7 *Presicion*, *Recall*, *F-measure* dan Akurasi dari Pengujian K-NN

K-values	*Precision*	*Recall*	*F-Measure*	Akurasi
2	1	0,86	0,924731	86%
4	1	0,62	0,765432	62%
6	1	0,62	0,765432	62%
8	1	0,64	0,780488	64%
10	1	0,6	0,75	60%
15	1	0,66	0,795181	66%
20	1	0,6	0,75	60%
25	1	0,56	0,717949	56%
30	1	0,56	0,717949	56%
35	1	0,54	0,701299	54%
40	1	0,58	0,734177	58%
45	1	0,52	0,684211	52%
50	1	0,5	0,666667	50%
75	1	0,54	0,701299	54%
100	1	0,28	0,4375	28%

Tabel 5.7 dapat diambil kesimpulan bahwa hasil yang dihasilkan oleh metode Improved K-NN lebih baik jika dibandingkan dengan hasil yang didapatkan dari metode K-NN.



Gambar 5.6 Grafik *Precision*, *Recall*, dan *F-Measure* dari Pengujian K-NN

Gambar 5.6 menunjukan grafik dari hasil pengujian menggunakan K-NN, garis berwarna biru mewakili nilai *precision*, garis berwarna merah mewakili nilai *recall*, dan garis berwarna hijau mewakili nilai *f-measure*. Grafik menunjukan bahwa pada nilai *k*=2 mengalami peningkatan nilai *f-measure* sedangkan pada nilai *k=100* nilai *f-measure* mengalami penurunan. Hal ini menunjukan bahwa pada nilai *k* yang rendah maka nilai *f-measure* akan semakin tinggi sedangkan semakin besar nilai *k* maka nilai *f-measure* akan semakin rendah, ini membuktikan bahwa hasil yang didapatkan dengan menggunakan metode K-NN memiliki tingkat akurasi yang rendah, serta penentuan nilai *k* merupakan hal yang perlu diperhatikan karena sangat berpengaruh pada tingkat akurasi.

Agar perbandingan hasil pengujian lebih jelas maka ditunjukan perbandingan antara hasil pengujian dari Improved K-NN dengan menggunakan skenario 5 yang merupakan skenario terbaik dengan hasil pengujian dari K-NN. Pada Tabel 5.8 menunjukan perbandingan dari hasil pengujian antara Improved K-NN dengan skenario 5 dan K-NN.

Tabel 5.8 Perbandingan Hasil Pengujian Improved K-NN Skenario 5 dan KNN.

k-values	Improved K-NN				K-NN			
	Precision	***Recall***	***F-Measure***	**Akurasi**	***Precision***	***Recall***	***F-Measure***	**Akurasi**
2	1	0,54	0,701298701	54%	1	0,86	0,924731	86%
4	1	0,7	0,823529412	70%	1	0,62	0,765432	62%
6	1	0,58	0,734177215	58%	1	0,62	0,765432	62%
8	1	0,66	0,795180723	66%	1	0,64	0,780488	64%

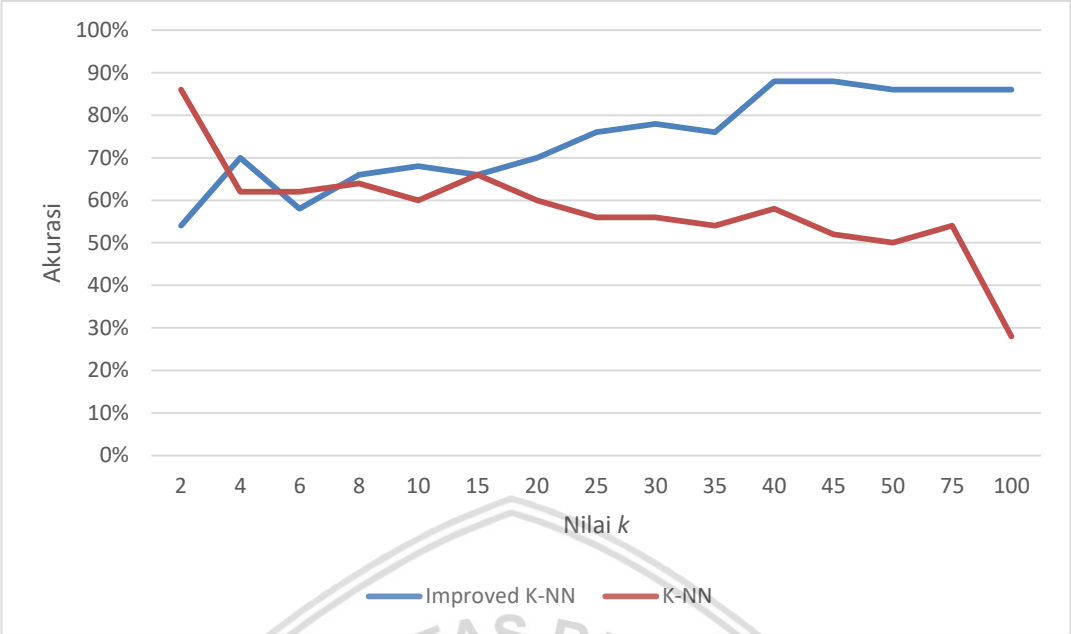
10	1	0,68	0,80952381	68%	1	0,6	0,75	60%
15	1	0,66	0,795180723	66%	1	0,66	0,795181	66%
20	1	0,7	0,823529412	70%	1	0,6	0,75	60%
25	1	0,76	0,863636364	76%	1	0,56	0,717949	56%
30	1	0,78	0,876404494	78%	1	0,56	0,717949	56%
35	1	0,76	0,863636364	76%	1	0,54	0,701299	54%
40	1	0,88	0,936170213	88%	1	0,58	0,734177	58%
45	1	0,88	0,936170213	88%	1	0,52	0,684211	52%
50	1	0,86	0,924731183	86%	1	0,5	0,666667	50%
75	1	0,86	0,924731183	86%	1	0,54	0,701299	54%
100	1	0,86	0,924731183	86%	1	0,28	0,4375	28%

Tabel 5.8 menunjukan hasil dari perbandingan masing-masing nilai *precision, recall, f-measure* dan akurasi dari pengujian dengan menggunakan skenario 5 pada Improved K-NN dan K-NN. Dilakukan perhitungan kembali dari *k-values* awal sehingga menjadi n (*k-values* baru) untuk setiap kategori yang dihitung menggunakan persamaan (2.6).

Hasil pengujian dari Improved K-NN dengan menggunakan skenario 5 menunjukan, nilai *f-measure* tertinggi ada pada *k-values* awal dengan nilai 40 dan 45 yakni sebesar 0,936170213 dan terendah ada pada *k-values* awal dengan nilai 2 yakni sebesar 0,701298701. Sedangkan pada hasil pengujian dari K-NN dengan menggunakan skenario 5 menunjukan, nilai *f-measure* tertinggi ada pada *k-values* awal dengan nilai 2 yakni sebesar 0,924731 dan terendah ada pada *k-values* awal 100 yakni sebesar 0,4375.

Hal tersebut menunjukan bahwa, pada metode K-NN semakin kecil *k-values* awal maka semakin tinggi nilai *f-measure* dan akurasi yang didapatkan. Terbukti pada *k-values* awal 2 tingkat akurasi pada metode K-NN sebesar 86% dan pada *k-values* 100 tingkat akurasinya sebesar 28%. Sedangkan metode Improved K-NN pada nilai *f-measure* dan akurasi untuk *k-values* awal 2 hasil yang didapatkan merupakan hasil yang terendah, dimana tingkat akurasinya sebesar 54% dan pada *k-values* 100 tingkat akurasi meningkat menjadi 76%. Selain itu pada pengujian dengan menggunakan metode Improved K-NN tingkat akurasi terbesar berhasil didapatkan pada *k-values* 40 dan 45 yakni sebesar 86%. Pada metode Improved K-NN tingkat akurasi yang didapatkan untuk masing-masing *k-values* tidak ada yang dibawah 50% sedangkan pada metode K-NN terdapat akurasi dibawah 50%.

Hal ini membuktikan bahwa hasil yang dihasilkan oleh metode Improved K-NN lebih baik dibandingkan hasil dari metode K-NN, serta metode Improved K-NN lebih stabil dibandingkan dengan metode K-NN terbukti dari hasil yang ditunjukan oleh Tabel 5.8.



Gambar 5.7 Grafik Perbandingan Improved K-NN dan K-NN

Gambar 5.7 menunjukan grafik dari perbandingan hasil pengujian menggunakan Improved K-NN dan K-NN, garis berwarna biru mewakili Improved K-NN dan garis berwarna merah mewakili K-NN. Grafik menunjukan bahwa pada nilai *k*=2 pada Improved K-NN tingkat akurasinya berada pada 54% sedangkan pada K-NN tingkat akurasinya berada pada 86%. Pada Improved K-NN untuk setiap nilai *k* tingkat akurasinya mengalami kenaikan sedangkan pada K-NN semakin tinggi nilai *k* semakin rendah pula akurasi yang didapatkan.

Pada Improved K-NN akurasi tertinggi berada pada nilai 88% pada nilai *k* = 40 dan *k* = 45, sedangkan pada K-NN akurasi tertinggi berada pada nilai 86 persen pada nilai *k* = 2. Selain itu untuk nilai akurasi terendah pada Improved K-NN berada pada nilai 54% pada nilai *k*=2, sedangkan pada K-NN nilai akurasi terendah berada pada nilai 28% pada nilai *k*=100. Hal ini membuktikan bahwa hasil yang didapatkan dengan menggunakan Improved K-NN lebih baik dibanding dengan K-NN, karena tingkat akurasi Improved K-NN tidak berada dibawah 50% sedangkan K-NN memiliki tingkat akurasi yang rendah yaitu berada dibawah 30%.

5.2 Analisis

Dari hasil pengujian yang dilakukan untuk setiap skenario pengujian yang digunakan, dapat diketahui bahwa beberapa faktor dapat memberikan pengaruh terhadap keakuratan hasil klasifikasi yang dilakukan dengan menggunakan metode Improved K-NN. Berdasarkan evaluasi yang telah dilakukan dengan menggunakan data uji sebanyak 50 data uji, dapat diketahui bahwa semakin banyak jumlah data latih yang digunakan maka akan semakin baik pula nilai *f-measure* yang dihasilkan. Skenario 1 hingga skenario 5 menunjukan peningkatan rata-rata nilai *f-measure.*

Rata-rata nilai *f-measure* yang paling rendah terdapat pada skenario 1, yang disebabkan karena data latih yang digunakan pada skenario 1 memiliki jumlah yang paling sedikit, serta perbandingan data latih untuk tiap kategori pada skenario 1 menggunakan data latih yang paling sedikit untuk setiap kategorinya.

Nilai *precision*, *recall,* dan *f-measure* yang paling rendah didapatkan apabila *k-values* awal yang digunakan terlalu kecil misal 2 atau terlalu banyak seperti yang ditunjukan oleh hasil setiap skenario yang mengakibatkan terjadinya kesalahan pada hasil pengklasifikasian. Hal ini membuktikan bahwa diperlukan ketelitian dalam menentukan *k-values* awal yang terbaik sehingga dapat menghasilkan hasil kategori yang tepat.

BAB 6 PENUTUP

6.1 Kesimpulan

Dari hasil penelitian sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN yang telah dilakukan dapat ditarik kesimpulan sebagai berikut :

1. Metode *Improved K-Nearest Neighbor* dapat dimanfaatkan dalam proses pengklasifikasian film dengan masukan berupa sinopsis film. Dokumen berupa sinopsis film akan melewati beberapa proses yakni *preprocessing*, pembobotaan *term*, hingga perhitungan nilai *cosine similarity* pada data latih yang digunakan. Kemudian proses selanjutnya ialah dengan mengurutkan tingkat kemiripan, menentukan *k-values* yang baru hingga mendapatkan hasil klasifikasi berupa kategori terhadap dokumen.
2. Dari hasil pengujian pada penelitian sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN didapatkan hasil terbaik yakni *precision* sebesar 1, *recall* sebesar 0,88, *f-measure* sebesar 0,936170213 dan akurasi sebesar 88%. Yang mana jumlah dokumen, perbandingan data latih serta *k-values* yang digunakan memiliki pengaruh atas baik atau tidak baiknya peoses pengklasifikasian pada dokumen yang berupa sinopsis.
3. Skenario pengujian yang digunakan untuk membandingkan hasil pengujian pada metode Improved K-NN dengan metode K-NN menggunakan data latih sebanyak 200 dokumen. Dari hasil pengujian yang telah dilakukan maka dapat ditarik kesimpulan yakni metode Improved K-NN dapat menghasilkan hasil yang lebih baik dengan hasil akurasi rata-rata sebesar 76% sedangkan jika dibandingkan dengan metode K-NN hanya menghasilkan hasil akurasi sebesar 54%.

6.2 Saran

Dari hasil penelitian sistem Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan Improved K-NN didapatkan beberapa saran agar dapat dikembangan lebih lanjut ialah sebagai berikut :

1. Pengklasifikasian yang dilakukan bergantung pada proses perhitungan *cosine similarity* yang berdasar pada frekuensi kemunculan kata. Melakukan pengecekan sinonim pada kata serta kemiripan kata berlandaskan makna katanya akan membantu untuk mendapatkan hasil yang lebih optimal.
2. Sistem dibangun dengan memanfaatkan metode Improved K-NN yang kurang mampu menangani jumlah data latih yang kurang seimbang secara tepat. Pengembangan sistem dengan menggunakan metode yang lain atau menggunakan metode Improved K-NN yang digabungkan dengan metode lain akan mampu memberikan hasil pengklasifikasin yang lebih optimal.

DAFTAR PUSTAKA

Sremanthy, J., & Balamurugan, P.S. (2012). *An efficient text classification using knn and naive bayesian*. International Journal on Computer Science and Engineering (IJCSE). Coimbatore, India.

Megantara, G., Kurniati, A.P., & Suryani, A.A., (2010). *Klasifikasi teks dengan menggunakan improved k-nearest neighbor algorithm*. Teknik Informatika, Fakultas Informatika, Universitas Telkom, Bandung.

Puspitasari, A.A., Santoso, E., & Indriati. (2018). *Klasifikasi dokumen tumbuhan obat menggunakan metode improved k-nearest neighbor*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 2, No. 2, Oktober 2018, hlm. 3948-3956. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Nathania, D.Z., Indriati., & Bachtiar, F.A. (2018). *Klasifikasi spam pada twitter menggunakan metode improved k-nearest neighbor*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 2, No. 10, Oktober 2018, hlm. 3948-3956. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Feldman, R., & Sanger, J. (2007). *The text mining handbook advance approaches in analyzing unstructured data* . CAMBRIDGE UNIVERSITY PRESS.

Wahyudi, D., Susyanto, T., & Nugroho, D. (2013). *Implementasi dan analisis algoritma stemming nazief & adriani dan porter pada dokumen berbahasa indonesia*. Jurnal Ilmiah SINUS STMIK Sinar Nusantara Surakarta. Program Studi Teknik Informatika, STMIK Nusantara Surakarta, Surakarta.

Xia, T., & Chai, Y. (2011). *An improvement to tf-idf: term distribution based term weight algorithm*. Journal of Software, 6(3), pp.413–420.

Manning, C.D., Raghavan, P., & Schutze, H. (2009). *An introduction to information retrieval*. Cambridge, England: Cambridge University Press.

Bagaskoro, G.N., Fauzi, M.A., & Adikara, P.P. (2018). *Penerapan klasifikasi tweets pada berita twitter menggunakan metode k-nearest neighbor dan query expansion berbasis distributional semantic*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 2, No. 10, Oktober 2018, hlm. 3948-3956. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Zheng, W., Wang, H., Ma, L., & Wang, R. (2015). *An improved k-nearest neighbor classification algorithm using shared nearest neighbor similarity*. 26(10), pp.133–137.

Prayoga, F., Pinandito, A., & Perdana, R. (2017). *Rancang bangun aplikasi deteksi spam twitter menggunakan metode naive bayes dan knn pada perangkat bergerak android*. Jurnal Pengembangan Teknologi Informasi dan Ilmu

Komputer, vol. 2, no. 2, p. 554-564, agu. 2017. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Herdiawan. (2015). *Analisis sentimen terhadap telkom indihome berdasarkan opini publik menggunakan metode improved k-nearest neighbor*.

Baoli, L., Shiwen, Y., & Qin, L. (2003). *An improved k-nearest neighbor algorithm for text categorization*. *Reading*, p.678.

Ting K.M. (2017). *Confusion matrix. In: sammut c., webb g.i. (eds) encyclopedia of machine learning and data mining*. Springer, Boston, MA.